Online ISSN : 0975-4172
Print ISSN : 0975-4350

# Global Journal of Computer Science and Technology: C

## Software & Data Engineering

Topic-Term Modeling

Research of Data Mining

Highlights

Hadoop Architecture

Weighted Associative Classifier

Discovering Thoughts, Inventing Future

VOLUME 14 ISSUE 2 VERSION 1.0

# Global Journal of Computer Science and Technology: C
# Software & Data Engineering

# Global Journal of Computer Science and Technology: C
## Software & Data Engineering

Volume 14 Issue 2 (Ver. 1.0)

Open Association of Research Society

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://globaljournals.us/terms-and-condition/menu-id-1463/

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**
**Employer Identification Number (EIN):**
**USA Tax ID: 98-0673427**

Global Journals Inc.

*(A Delaware USA Incorporation with "Good Standing"; **Reg. Number: 0423089**)*

*Sponsors: Open Association of Research Society*
*Open Scientific Standards*

## *Publisher's Headquarters office*

Global Journals Headquarters
301st Edgewater Place Suite, 100 Edgewater Dr.-Pl,
**Wakefield MASSACHUSETTS,** Pin: 01880,
United States of America
*USA Toll Free: +001-888-839-7392*
*USA Toll Free Fax: +001-888-839-7392*

## *Offset Typesetting*

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

## *Packaging & Continental Dispatching*

Global Journals
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin:452009, 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: *investors@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) & 250 USD (Color)

## Integrated Editorial Board
## (Computer Science, Engineering, Medical, Management, Natural Science, Social Science)

## President Editor (HON.)

## Chief Author (HON.)

## Dean & Editor-in-Chief (HON.)

# Contents of the Volume

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Research of Data Mining Algorithm based on Cloud Database

By Tianxiang Zhu, Xia Zhang, Dan Zhang & Xin Liu

*Shenyang University of Technology, China*

*Abstract-* There is an immense amount of data in the cloud database and among these data, much potential and valuable knowledge are implicit. The key point is to discover and pick out the useful knowledge, and to do so automatically. In this paper, the data model of the cloud database is analyzed. Through analyzing and classifying, the common features of the data are extracted to form a feature data set. The relationships among different areas in the data are then analyzed, from which the new knowledge can be found. In the paper, the basic data mining model based on the cloud database is defined, and the discovery algorithm is presented.

*Keywords: cloud database, data mining, association rules, classification characteristic.*

*GJCST-C Classification: H.2.8*

# Research of Data Mining Algorithm based on Cloud Database

Tianxiang Zhu [α], Xia Zhang [σ], Dan Zhang [ρ] & Xin Liu [ω]

*Abstract-* There is an immense amount of data in the cloud database and among these data, much potential and valuable knowledge are implicit. The key point is to discover and pick out the useful knowledge, and to do so automatically. In this paper, the data model of the cloud database is analyzed. Through analyzing and classifying, the common features of the data are extracted to form a feature data set. The relationships among different areas in the data are then analyzed, from which the new knowledge can be found. In the paper, the basic data mining model based on the cloud database is defined, and the discovery algorithm is presented.

*Keywords: cloud database, data mining, association rules, classification characteristic.*

## I. Introduction

Cloud computing is derived from technologies such as distributed processing, parallel processing, grid computing, etc. It is an emerging approach to sharing the infrastructure architecture[1]. It distributes all the computing tasks on the resource pool that is made of many computers, making sure all the application systems can acquire desired computing power, memory space and software service according to their demand[2]. All the computing is provided to the terminal user by the form of service, and all the application software in the cloud as shared resources. A cloud database is a database deployed and virtualized in the cloud computing environment. It is predicated that as it develops overtime, more and more people and companies will store all their data in the cloud, which will make data mining based on the cloud computing one of the trends in the future data mining systems[3].

There is a massive amount of data in the cloud database, and among them, lives potentially valuable knowledge. How to discover such useful knowledge is the key point in database research. Data mining is the process of picking out the hidden knowledge and regulations, which possess potential value that could influence decision making[4]. Data mining namely refers to the knowledge discovery from a database and is comprised of the following procedures: data pre-processing, data alternating, data mining operation, rule expression and evaluation[5]. A data mining system includes: control unit - used to control all parts in a harmonious way; database interface – used to generate and process data according to the given query; database - used to store and manage relevant knowledge; focus - refers to the data extent that needs to be inquired; model extracting - refers to the various data mining algorithms; and finally, knowledge evaluation- used to evaluate the extracted conclusion[6].

## II. Cloud Database

A distributed database is a logical set of the databases at various sites or nodes in a computer network and logically, such databases belong to the same system[7]. Different from the traditional distributed database, a cloud database contains isolated as well as shared data; a cloud database can be designed by using different data models, which mainly include the key-value model and relationship model.

All data of the key-value model, including the rows and columns, are stored in the cells of a table. Contents are partitioned by row, the rows make up a tablet, and the tablet is stored on a server node.

### *a) Row Key*

Data is maintained in the lexicographic order on the row key. For a table, a row interval is dynamically partitioned according to the value of the row key and is the basic unit in which load balancing and data distribution are performed. Row keys are distributed amongst data servers.

### *b) Column Key*

Column keys are grouped into sets of many “column families” and are the basic units in which access control is performed. All data stored in a column family usually belong to the same data type, which means data is compressed at a higher rate. Data can be stored in a column key of the column family.

### *c) Timestamp*

Each cell contains multiple versions of the same data and these versions are indexed by the timestamp. Data model for key-value cloud database is as shown in Fig. 1:

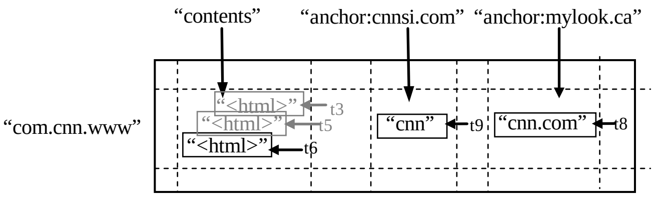


*Figure 1 :* Data model for key-value cloud database

*Author α σ ρ : School of Software, Shenyang University of Technology Shenyang, Liaoning, China. e-mail: zhutianxiang@gmail.com*
*Author ω : Liaoning Northern Laboratory CO., LTD, Shenyang, Liaoning, China.*

The data model for the relational cloud database involves such relevant terms as row group and table group. A table is a logical relationship and includes a partitioning key, which is used for partitioning the table. The set of many tables with the same partitioning key is called a table group. In that table group, the set of rows with the same partitioning key value is called a row group. The rows in that row group are always allocated to the same data node. Each table group contains many row groups, which are allocated to different data nodes. A data partition contains many row groups, so each data node stores all rows with a certain partitioning key value. The data model for the relational cloud database is as shown in Fig. 2:

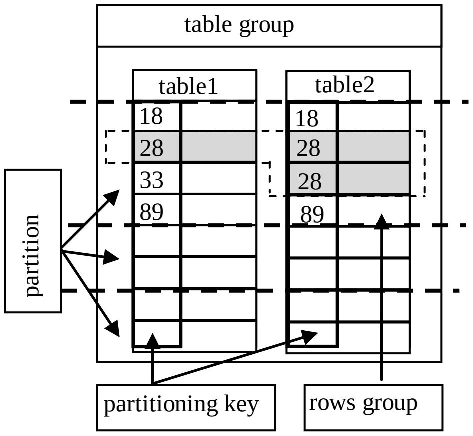


*Figure 2 :* Data model for relational cloud database

## III. Data Mining for Association Rules

### *a) Model of the Association Rules*

The normal target of the association rules is to discover the data relations among the data item set in the relationship type cloud database. Through mining based on the association rules, we can discover the relevance of the data.

In the subject item set, there are some target features in the relationship type cloud database. For instance, the commodity data item set in the commercial behavior analysis {T-shirt, coat, shoes, milk, bread ... }; data item set in the medical diagnosis analysis {hypertension, diabetes... }.

Classifying item set has the similar features with the subject item set, for instance, customer data item set in the commercial behavior analysis {vocation, gender, age... }; diagnosis behavior in medical diagnosis and signs and symptoms item set {smoking, polysaccharide, hyperlipidemia ... }.

Sample item set, which has both the features in the subject item set and the transaction data item set in the classifying item set. For instance, transaction data in the commercial activity analysis { {Zhangsan, milk}, {Zhangsan, bread}, {Lisi, T-shirt} ... }, health check information in the medical diagnosis {{ Zhangsan, smoking, hypertension}, {Lisi, hyperlipidemia, diabetes}... }.

Through the mining based on the association rules, we can find that 90% of the customers who bought milk also bought bread; 50% of the patients who have hyperlipidemia also have diabetes.

The common targets of the association rules are transaction databases with the characters of subjects oriented item set. In practice, most databases are relational, and many applications and the required knowledge are from many different item sets（or multi-item set for simplicity）. For relational databases, it is difficult to describe the complicated association rules between the multi-item set with models of general association rules. We present the association rules model of multi-item set for the relational databases:

*Definition 1:* I is the subject item set, J is the taxonomy item set, each transaction corresponds to a subset T of the subject item sets and a taxonomy item U of the taxonomy item sets, called T belonging to class U.

*Model 1:* it is supposed that R=(r1,r2,…, rn) is the rows group in the relational cloud database, rk is one of the rows item set, D is a sample item set relevant to R, and each sample d corresponds to one rows item set, i.e. $d\subseteq R$. Each sample is marked with SID (sample identifier). As for the classifying item set X, only when $X\subseteq d$, the sample X belongs to d. association rules is a formula like $X\subseteq d\Rightarrow Y\subseteq d$, it can be $X\Rightarrow Y$, therein, $X\subseteq R$, $Y\subseteq R$ and $X\cap Y=\Phi$.

The rule $X\Rightarrow Y$ in the sample item set D is constrained by degree of confidence C and degree of support S. Degree of confidence C is defined as C% in the transaction X in D also contains Y. Degree of support S is defined as transaction $X\cup Y$ accounts for S% in D. Degree of confidence represents the strength of the rule, while Degree of support means the frequency of the model, which is shown in the rule.

In the cloud database containing cases information, 66% of the crime site in the theft cases happened in factories, so the C is 66%. Theft cases and factory cases account for 17% of the total cases, so the S is 17%.

The data frequency item set can be defined as the data item set where the degree of support S is over the pre-defined minimum degree of support S. The association rules with high degree of support S and degree of confidence C is considered strong association rules, otherwise it is considered weak association rules. Association rules mining means to find the line group that accord to the strong association rules in the database.

The procedure for mining these kinds of association rules of multi-item set is as follows:

1. Divide transaction D into several transaction subset D'={D1',D2',…Dn'} according to taxonomy item sets.
2. For all D1'<D' Do
   Find the strong sets of the main subject item
   Derive the association rules using the strong set
3. Next

These association rules of the multi-item set possess a feature where only one value is available in each sample (SID) set. With this method, mining the data's association rules is applicable for one-to-many relational databases. This is more practical and expands the mining range for the association rules.

In practice, most of the applications and knowledge is from the multiple data item set. For example, we regard a criminal case as the sample item set. For each case, there is one mark SID, several suspects, as well as several methods by which the crime is committed. So we can first take the education level of the suspects as one data item set, and the methods of committing crime as another data item set.

There are association rules with several multi-item sets, the association rules model can be termed as:

*Model 2:* It is supposed that $I=(i_1,i_2,\dots, i_n)$ is a classifying item set, $J=(j_1,j_2,\dots,j_m)$ is another one, D is a sample item set, each sample has two classifying item sets $T(T\subseteq I)$ and $T'(T'\subseteq J)$, and each sample is marked with SID. The formula is $X\subseteq I\Rightarrow Y\subseteq J$, degree of confidence C can be termed as that in sample where D contains $X\subseteq I$, C% has $Y\subseteq J$, degree of support S can be defined as transaction with $X\subseteq I$ and $Y\subseteq J$ accounts for S% in D.

### b) *Mining Algorithm*

There are many algorithms in the association rules, and the representative Apriori Algorithm follows the rule that the sub-item sets of all the strong item sets are classified to the strong item sets, while the super item sets of the weak item sets are weak item sets.

The first pass of the algorithm simply counts item occurrences to determine the strong 1-itemsets. A subsequent pass, pass k, consists of two phases. First, the strong item sets L found in the (k-1)th pass are used to generate the candidate item sets Ck, using the apriori-gen function. Next, the database is scanned and the support of candidates in Ck is counted. For fast counting, we need to efficiently determine the candidates in Ck that are contained in a given samples.

As for the association rules of multiple data item sets, we need to have strong item sets L1 with item 1, and then we can have C2 from L1 with the item 2, after this we can have L2, based on this method we can finally have Ck, and get Lk from the database.

Classifying item set D into m classifying item sets D1, D2, ... Dm according to the separating item set J, then we can find out the association rules after using Apriori Algorithm to each sub-sample item set D.

```
for(j=1;j<=m;j++) do
 begin
   $L_{j,1}$={large 1-items};
     for (k=2;$L_{j,k-1} \neq \Phi$;k++) do
       begin
         $C_k$=apriori-gen($L_{j,k-1}$);
         forall samples s$\in D_j$ do
           begin
             Cs=subset($C_k$,s);
             forall candidates c$\in$Cs do
              c.count++;
           end
         $L_{j,k}=\{c\in C_k \mid c.count>=minsup\}$
       end
     Answer=$\cup_{j,k} L_{j,k}$;
   end;
```

Lj,1 represents the strong item set in Dj sub sample item set, which will generate K item in Dj, scan the database to have Lj,k , we finally can have D1,D2,…,Dm strong item set from the sub sample item set.

Since Model 2 corresponds to two classifying item sets and each sample $S\subseteq D$ includes classifying item set I and J, 1-itemsets represent the strong item sets we select from I and J, which is Li,j. From Li,j we can have C1,2 from L1,2, done with the similar manner, and then get L1,k. From L1,1, we can have C2,1 from L2,1, the algorithm is:

```
$L_{1,1}$={1-itemsets x$\in$I, y$\in$J};
 for (i=2;i=n;i++) do
begin
 for(j=2;j<=m;j++) do
   begin
     $C_{i,j}$=apriori-gen($L_{i,j-1}$);
     forall samples s$\in$D do
       begin
         Cs=subset($C_{i,j}$,s);
         forall candidates c$\in$Cs do
          c.count++;
     end
   $L_{i,j}=\{c\in C_{i,j} \mid c.count>=minsup\}$
  end
  Answer=$\cup_{i,j} L_{i,j}$;
End
```

In management information systems, the relational database is widely used; the connection among different data is one-to-many and many-to-many, so it is universal to discover knowledge in the database. As the cloud age is coming, data mining from the cloud data is more practical. The mining method that is used in the association rules is applied to the cloud database, making the association rules more practical and universal. This paper also extends the Apriori Algorithm into association rules mining model, which realize the mining multi-item set association rules.

## IV. Data Mining for Classification Characteristic Rule

Knowledge discovered from a database with massive data is diversified in variety. Knowledge classification refers to clustering or classifying tuples in the database to divide these tuples into different categories by characteristic rules extracted from a certain target class, and thus achieve the purpose of describing the characteristics of the tuples of that class.

Clustering refers to categorizing a group of individuals into several categories, which means those with the same characteristic are classified as one category. Clustering is a process in which a data object with multiple attributes is continuously classified. In such process, classification is automatically executed by the classification algorithm to divide the data into several classes by identifying data features. A relational database mainly containing character information may be equivalently partitioned into equivalence classes according to the concept of equivalence class. The resulting equivalence classes are a group of classes. The characteristics of each class are further analyzed and this can lead to the determination of the classification characteristic rules. Such analysis process is of practical significance. For example, symptoms and reaction characteristics of various diseases can be determined by analyzing a great amount of medical diagnosis cases.

### *a) Classification Model For Key-Value Model Based Cloud Database*

Let D be a key-value model based cloud database, K represents the set of all row keys in D with the formula K={k1, k2 ,…,kn}, At represents the set of all column keys in D with the formula A ={a1,a2,…,am}, V represents the dataset of certain attribute characters of the column keys with the formula V={v11,v12,…,vmn} and f represents a function of a and k with the formula Vi,j=f(ai,kj).

*Definition 2:* For $\forall a \in A_t$ ($A_t$ is the dataset of column keys, $A_t \subseteq A$), if $k_i \in K$, $k_j \in K$, $i \neq j$ and f(a, ki)= f(a, kj), then ki is said to be equivalent to kj based on the dataset of column key attributes At, and the set of all equivalent row keys based A is called equivalence class based on the dataset of column key attributes A; all row sets in K are classified by equivalence class and the classification result is called A-based classification: K={K1, K2, …}.

*Definition 3:* K={K1, K2, …}, K is the At-based classification and a column key in At is called a classification. The attribute value of the column key in At is called the name of the classification.

*Definition 4:* Let D be a key-value model based cloud database, $S_k$ represents the amount of the latest timestamps in the row key set, At is a column key set of D, Y is a At-based equivalence class and $S_y$ is the amount of the latest timestamps in the set of row keys in Y, then $S=S_y/S_k*100\%$ is said to be the classification support degree of the equivalence class Y.

*Example 1:* Let D be a key-value model based cloud database, K is the set of all row keys in D, A is the set of all column keys in D and At is a subset of A. V is the dataset of certain attribute characters of the column keys and each data has the latest timestamp.

K={ k1, k2, k3, k4, k5, k6, k7, k8, k9, k10, k11, k12 }
A ={a1,a2,a3}
At ={a1}
V={ v10,v11, v12,v20, v21,v30, v31,v32}

The values of f(k,a) are as shown in Table 1.

*Table 1 :* The values of f(k,a)

| A<br>K | A1 | A2 | A3 |
|---|---|---|---|
| $k_1$ | $v_{11}$ | $v_{20}$ | $v_{32}$ |
| $k_2$ | $v_{10}$ | $v_{21}$ | $v_{30}$ |
| $k_3$ | $v_{12}$ | $v_{20}$ | $v_{30}$ |
| $k_4$ | $v_{11}$ | $v_{21}$ | $v_{30}$ |
| $k_5$ | $v_{11}$ | $v_{20}$ | $v_{32}$ |
| $k_6$ | $v_{12}$ | $v_{20}$ | $v_{30}$ |
| $k_7$ | $v_{10}$ | $v_{21}$ | $v_{31}$ |
| $k_8$ | $v_{11}$ | $v_{21}$ | $v_{31}$ |
| $k_9$ | $v_{11}$ | $v_{20}$ | $v_{32}$ |
| $k_{10}$ | $v_{10}$ | $v_{21}$ | $v_{31}$ |
| $k_{11}$ | $v_{11}$ | $v_{20}$ | $v_{32}$ |
| $k_{12}$ | $v_{10}$ | $v_{21}$ | $v_{31}$ |

In the above mentioned database, the field {a1} in the column key set At can be classified into three classes:

K1 = {k1, k4, k5, k8, k9, k11} ; K2 = {k2,k7,k10,k12} ; K3={k3,k6}. {K1,K2,K3} is a class based on the column key set At, the name of that class is {v11, v10, v12} and its classification support degree is {50%, 33.33%, 16.67%}.

### *b) Classification Model for Relational Model Based Cloud Database*

Let D be a relational model based cloud database and T be a table group of D, P represents the set of partitioning keys in T with the formula P={ p1, p2,…,pn } and R represents the set of row groups of the partitioning key Pi with the formula R={r1,r2,…,rn}.

*Definition 5:* R represents the set of row groups of the partitioning key Pi with the formula of R={r1, r2,…,rn } and is called a class based on the partitioning key P={ p1, p2,…,pn }. pn is called the name of the class rn.

*Definition 6:* Let D be a relational model based cloud database and T be a table group of D, $S_t$ represents the record count of all row groups, the row group set R is the class based on the partitioning key P and $S_y$ represents the record count of Y row groups in R, then $S=S_y/S_t*100\%$ is said to be the classification support degree of the class Y.

*Example 2:* Let D be a relational model based cloud database and T be a table group of D, P is the partitioning key and the value of P is {p1,p2,p3}, then the corresponding row group is {r1,r2,r3}, namely:

P ={p1,p2,p3}<br>
R={r1,r2,r3}<br>
r1={v11,v12,v13,v14,v15,v16}<br>
r2={v21,v22,v23}<br>
r3={v31}

The above mentioned database can be partitioned into three classes based on the partitioning key P:

r1={v11, v12, v13, v14, v15, v16}, r2={v21,v22,v23}, r4={v31}, the name of the class is {r1,r2,r3} and the classification support degree is {60%, 30%, 10%}.

For the cloud database D, all classification support degrees {S1,S2,S3,…} can be obtained according to a certain classification R={R1,R2, R3,…}.

*Definition 7:* Let Sp be a given threshold, $0≤S_p≤1$. Those classes with a classification support degree $S≥S_p$ are called strong class and those with a classification support degree $S<S_p$ are called weak classes.

In mining knowledge from massive data, we usually are concerned about and interested in data classes with higher classification support degree, namely the strong classes. Strong classes contain more representative knowledge.

### c) Classification Characteristic Rule Model

According to the definitions as mentioned above, data in a database can be classified and the characteristics of the strong classes need to be further analyzed.

*Definition 8:* Let E be a At-based class, $A_t$ is the complementary set of $A_t$ against the attribute A, $A_t⊆A$, B is the subset of $A_t$, the equivalence class T based on B is called the characteristic domain in the class E, and the value $\{b_1,b_2,…\}$ of the attribute in B is called the characteristics in the class E.

*Definition 9:* Let ec be the record count of the class E, and tc be the record count of the characteristic domain T, then $C= t_c/e_c*100\%$ is said to be the confidence degree of characteristic.

*Definition 10:* Let $C_p$ be a given threshold, $0≤C_p≤1$, a characteristic domain with the confidence degree of characteristic $C≥C_p$ is called a strong characteristic domain and a characteristic domain with the confidence degree of characteristic $C<C_p$ is called a weak characteristic domain. The value of the field with strong characteristic domain is called a strong characteristic, while the value of the field with weak characteristic domain is called a weak characteristic.

Strong characteristics in a strong class are usually representative knowledge and can be expressed as:

(E, T, Cp)<br>
E: class<br>
T: characteristic<br>
Cp: confidence degree of characteristic

Discovery Algorithm for Classification Characteristic Rules.

D is a cloud database and A is the set of classification attributes of D.

```
 For all $A_t⊆A$ Do
   Obtain the set of At-based classes $E_1, …,E_m$;
   Obtain the classification support degree of the class set $E_1, …,E_m$ $S=\{S_1, …,S_m\}$;
   For i=1 To m Do
      If $S_i ≥S_p$ Then
        Obtain all characteristic domains $T_1, T_2, …T_k$;
        Obtain the confidence degree of characteristic of the characteristic domain set $T=\{T_1, T_2, …T_k\}$ $C=\{C_1, C_2, …C_k\}$;
        For j=1 To k Do
          if $C_i ≥C_p$ Then
             $(E_i, T_j, C_j) ⇒$result base
          Endif
        Next
      Endif
   Next
Next
```

## V. Application of the Classification Characteristic Rules in Case Information Systems

Suppose the related property is the case type, selected site and the way of commit, the related degree of the door smashed versus picked is 0.8, the given threshold of the related degree is 2.5, two cases as shown in Table 2:

*Table 2 :* These Two Cases are Related

| | Case kind | Selected site | Way of commit |
|---|---|---|---|
| Case1 | Burglary | residence | door picked |
| Case2 | Burglary | residence | door smashed |

Based on the above definitions, as long as the related degrees of the related properties are known, the related cases can be discovered. The values of the related degrees are provided by the field experts according to the field knowledge. In order to express the related degrees, a related degree matrix Ma is defined as follows:

$$\text{Ma}=\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ C_{m1} & C_{m2} & \cdots & C_{mn} \end{bmatrix} \qquad (1)$$

Cij: related degree of element j to element i of property a.

$M_a$ is a symmetrical matrix, so only consider the lower triangle.

The related degree matrix of the way of commit is as shown in Table 3:

*Table 3 :* The Related Degree Matrix of the Way of Commit

| | Pick door lock | Tag along after | Smash door | Cut bag | Pretend to be conceal | Conceal |
|---|---|---|---|---|---|---|
| Prize door lock | 1 | 0 | 0 | 0 | 0 | 0 |
| Tag along after | 0 | 1 | 0 | 0 | 0 | 0 |
| Smash door | 0.8 | 0.5 | 1 | 0 | 0 | 0 |
| Cut bag | 0 | 0.9 | 0 | 1 | 0 | 0 |
| Pretend to be | 0 | 0.8 | 0.3 | 0 | 1 | 0 |
| Conceal | 0 | 0 | 0 | 0 | 0 | 1 |

During case analysis, the related degree matrix of all related properties must be known. The sum of case related degrees can be found based on the related properties and the related degree matrix and all the related cases can be found based on the given threshold Cp.

Input U, an information system, has n records, Ma is a symmetrical matrix, Mak[i,j] seeks the correlation degree in the correlation matrix.

Output L, a set of case related to it.

```
For i=1 To n-1 (n is the record number)
   For j=i+1 TO n
      A [i, j] =0
         For k=1 TO m
            A[i,j]=A[i,j]+Mak[i,j]
         Next
         If A [i,j]>=Cp
   uj∪{u | u relates to ui }
         End If
   Nexts
   L=U{u | u relates to ui }
Next
Output L
```

Following the idea of converging classes, the case information is divided into many kinds with the equivalent dividing method. Define the base value of kinds as classifying the support degrees. The kinds can be divided into strong class ones and weak class ones. The weak class has too small classifying support degrees, no practical meanings and can be neglected. For the strong class, Rough set theory can be used to analyze their common features and form the classifying characteristic regulations.

Data was mined from the database for the experimental criminal case information system by using the aforementioned algorithm. Taking the crime approach table group as an example, the table contains 100762 records. Given a classification support degree threshold of 10%, and a characteristic confidence degree threshold of 20%, 359 classification characteristic rules were mined, for example:

(Residential house, night, 23.4%)
(Rubbery, less than RMB10000, 93.3%)

## VI. Conclusions

The data mining technique is new to the information society. Many subjects need to be studied in this field. In many professions, a certain amount of databases have been accumulated, in which some hidden knowledge needs to be discovered. Starting with the concept of set theory, the data model for the cloud database was analyzed; the model and algorithm for mining classification characteristic rules from cloud database were designed to make data mining of classification characteristic rule more practical.

The abstracted related knowledge models presented in this paper can be put into practice in the public security affairs, such as case chaining, which is one of the highly demanded, complex tasks in the public security affairs. The presented methods about the related case data mining in this paper promote the work effect of the chained cases. On the case material analysis, the mining of the classifying characteristic regulations help users with their classifying work and overcome the weaknesses that exist in the old statistics method, in which repeated experimentation are required.

## References Références Referencias

1. Lin ZY, Lai YX, Lin C, Xie Y, Zou Q. Research on cloud databases. Journal of Software, 2012,23(5):124-137.
2. Xu M, Gao D, Deng C, Luo ZG, Sun SL. Cloud computing boosts business intelligence of

telecommunication industry. In: Jaatun MG, Zhao GS, Rong CM, eds. Proc. of the 1st Int'l Conf. on Cloud Computing (CloudCom 2009). Berlin: Springer-Verlag, 2009. 224−231.

3. Feng DG, Zhang M, Zhang Y, Xu Z. Study on cloud computing security. Journal of Software, 2011,22(1):71−83.
4. Zhu TX, W P, Zhang D. Application of the data mining technique in case information systems. ICCSEE 2012, 3(1):43-46.
5. Zhu TX, Li L, Xu ZW, Technology for Mining Classification-Characteristic Rules, Journal of Shenyang Polytechnic University, 1999.(x)x:22-24 .
6. He JJ, Ye CM, Wang XB, Huang ZX, Liu QL. Cloud Computing-Oriented Data Mining System Architecture, Application Researche of Computers, 2011,28(4):1372-1374.
7. Abouzeid A, Bajda-Pawlikowski K, Abadi DJ, Rasin A, Silberschatz A. HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads. PVLDB, 2009,2(1):922−933.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Stochastically Simulating the Effects of Requirements Creep on Software Development Risk Management

By P. K. Suri & Shilpa Rani

*Kurukshetra University, India*

*Abstract-* One of the major chronic problems in software development is the fact that application requirements are almost never stable and fixed. Creeping user requirements have been troublesome since the software industry began. Several empirical studies have reported that volatile requirements are a challenging factor in most information systems development projects. Software process simulation modeling has increasingly been used for a variety of issues during software development. The management of software development risks is one of them. This study presents an approach for simulating and analyzing the effect of Requirements Creep on certain software development risk management activities. The proposed algorithm is based on stochastic simulation and has been implemented using C.

*Keywords: requirements creep, requirement volatility, requirement management, stochastic simulation, software risk management.*

*GJCST-C Classification: K.6.3*

# Stochastically Simulating the Effects of Requirements Creep on Software Development Risk Management

P.K. Suri [α] & Shilpa Rani [σ]

*Abstract-* One of the major chronic problems in software development is the fact that application requirements are almost never stable and fixed. Creeping user requirements have been troublesome since the software industry began. Several empirical studies have reported that volatile requirements are a challenging factor in most information systems development projects. Software process simulation modeling has increasingly been used for a variety of issues during software development. The management of software development risks is one of them. This study presents an approach for simulating and analyzing the effect of Requirements Creep on certain software development risk management activities. The proposed algorithm is based on stochastic simulation and has been implemented using C.

*Keywords: requirements creep, requirement volatility, requirement management, stochastic simulation, software risk management.*

## I. Introduction

Software process simulation modeling is increasingly being used to address variety of issues from the strategic management of software development, to supporting process improvements, to software project management training. One of the proposed purposes for software process simulation is the management of software development risks, usually discussed within the category of project management [1]. There have been various (but quite a limited) studies which have used modeling and simulation for software development risk management for example: Madachy's Model [2], Houston's Model [3]. The present study also describes an approach for managing software development risks using simulation.

In the present work, implementation of a simulator has been done for modeling the effects of Requirements Creep on various risk management factors during software development using stochastic simulation.

This paper has been has been organized into various sections including the present one. An overview of software development risk factors has been provided in section II while 'requirements' as a major risk factor during software development have been discussed in section III, followed by potential effects of requirements creep(section IV). The proposed algorithm has been provided in section V, the results of which have been demonstrated and interpreted in section VI with the help of charts representing the relationships between various risk management factors.

## II. Risk Factors during Software Development

Top 10 software risk items identified by Boehm [4] for software development projects:

- Personnel shortfalls
- Unrealistic schedules and budgets
- Developing the wrong functions and properties
- Developing the wrong user interface
- Gold plating(adding more functionality/ features than is necessary)
- Continuing stream of requirements changes
- Shortfall in externally furnished components
- Shortfalls in externally performed tasks
- Real-Time performance shortfalls
- Straining computer-science capabilities

Jones [5] has presented the following three key software areas:

- Risks associated with inaccurate estimating and schedule planning
- Risks associated with incorrect and optimistic status reporting
- Risks associated with external pressures, which damage software projects.

Some investigators have even presented software development risks on the order of 150 or more. Twenty nine of these risk factors have been cited by Houston [3] as most important Software development risk factors.

## III. Requirements: A Major Software Development Risk Area

A requirement is the condition or capacity that a system that is being developed must satisfy [6]. Requirement management in general is mainly concerned with three tasks: Requirement Elicitation, Requirement Analysis and Requirement specification.

*Author α: Dean (R&D), Chairman & Professor (CSE/IT/MCA), H.C.T.C.M., Technical Campus, Kaithal, Haryana, India.*
*e-mail: pksurikuk@gmail.com*
*Author σ: Research Scholar, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, Haryana, India.*
*e-mail: shilparanikuk@gmail.com*

One of the major chronic problems in software development is the fact that application requirements are almost never stable and fixed. Creeping user requirements have been troublesome since the software industry began. Several empirical studies have reported that volatile requirements are a challenging factor in most information systems development projects [7], [8], [9], [10]. There is no quick, perfectly effective cure. Various factors have been considered to be behind the creeping user requirements [3], [7], [11], [12], [13], [14], from which the following have been modeled in the presented study:

- Excessive Schedule Pressure
- User-Practitioner Relationship Level (which accounts for the User's involvement level and Practitioner's level of knowledge).

This study demonstrates the use of stochastic simulation as a flexible vehicle for effectively assessing and managing risk by measuring the effect of requirements creep on various software risk management factors using stochastic simulation.

## IV. Potential Effects of Requirements Creep

The requirements creep level may be affected by the high schedule Pressure and User Practitioner Relationship level which in turn may affect the Defect generation rate, rework and job size [15]. The present algorithm simulates the effect of requirements creep by sampling the distribution of variables and continuously recalculating them after each run.

## V. Algorithm

| Symbol Used | Interpretation |
|---|---|
| CL | Creep Level |
| CCL | Cumulative Creep Level |
| INCREASE | Increase in Creep Level |
| CINCREASE | Cumulative Increase in Creep Level |
| SHPL | Schedule Pressure Level |
| CSHPL | Cumulative Schedule Pressure Level |
| IJS | Increased Job Size |
| CIJS | Cumulative Increased Job Size |
| UPRL | User-Practitioner Relationship Level |
| IncIJS | Increase in Job Size per unit rise in Creep Level |
| RWC | Rework Cost |
| CRWC | Cumulative Rework Cost |
| IncRWC | Increase in Rework Cost per unit rise in Creep Level |
| SRUNS | Number of Simulation Runs |

```
STEP 1:  Read Input data.
    [Read SRUNS and UPRL]
STEP 2:  Do the initialization:
[Set CL=0, CCL=0, INCREASE=0, CINCREASE=0, SHPL=0, CSHPL=0, UPRL=0, CUPRL=0, RWC=0, CRWC=0, IJS=0, CIJS=0, IncIJS=0, IncRWC=0]
STEP 3: Generate Schedule Pressure Level (from a random distribution)
STEP 4: CL=CL+SHPL-UPRL
 RUN=RUN+1
STEP 5:  If ((SHPL-UPRL)<CL) THEN {
INCREASE= CL- SHPL-UPRL
CINCREASE=CINCREASE+INCREASE
(Generate random values of IncRWC and IncIJS)
CIJS=CIJS+ INCREASE*IncIJS
CRWC= CRWC+INCREASE*IncRWC}
Compute defect generation percentage w.r.t. requirements creep level each time.
STEP 6: Compute Average Creep Level, Average Schedule Pressure Level, Average Rework Cost and Average Increase in Job Size.
STEP 7: Compute percentage of Defect Generation with respect to requirements creep level=((CCL-CSHPL)/CCL)*100
STEP 8:  Print the computed statistics.
STEP 9:  If RUN < SRUNS then go to STEP 3.
              (Run for a large value of SRUNS)
STEP 10: END.
```

## VI. Results & Interpretation

Table 1

| User-Practitioner Relationship Level | Average Requirements Creep Level | |
|---|---|---|
| | Avg. Schedule Pressure=5 | Avg. Schedule Pressure=8 |
| 1 | 220.607 | 343.166 |
| 2 | 173.958 | 296.061 |
| 3 | 128.358 | 250.334 |
| 4 | 83.534 | 204.735 |
| 5 | 38.884 | 159.708 |
| 6 | 6.5289 | 115.058 |
| 7 | 1.487 | 70.408 |
| . | . | . |
| . | . | . |

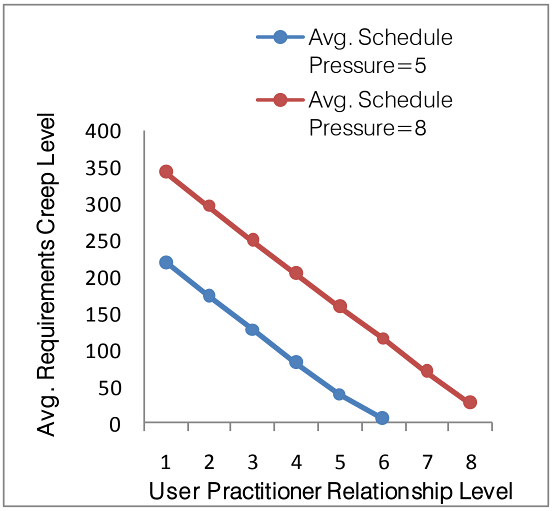


*Figure 1 :* Variation of Average Requirements Creep with User-Practitioner Relationship level

From the table1, it can be analyzed that a boost in User-Practitioner Relationship Level (i.e. User Involvement Level and Practitioner's Level of Knowledge) lowers the average level of Requirements Creep at a given schedule pressure.

Defect generation percentage calculated with repect to the level of rerquirement creep can be analysed with the help of figure2:

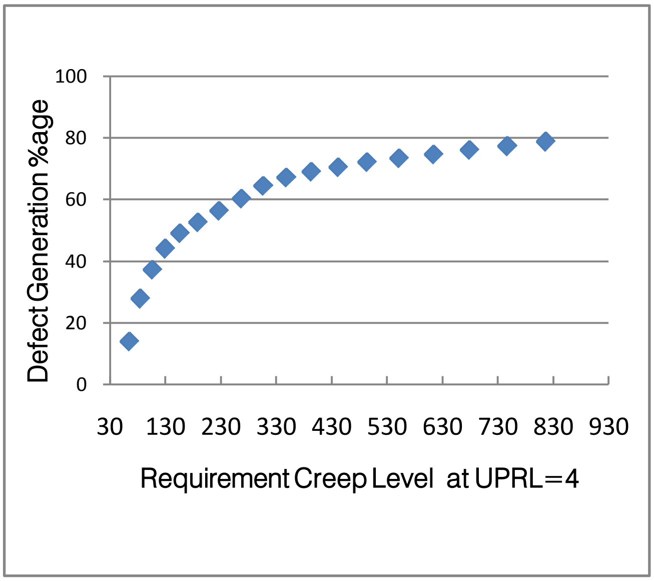


*Figure 2 :* Defect generation percentage with respect to requirements creep level at a certain User-Practitioner Relationship level

Increasing level of requirements creep may result into an increase in defect generation percentage.

| User -Practitioner Relationship Level | AVERAGE REWORK COST at | | | |
|---|---|---|---|---|
| | Avg. SHPL=5 | Avg. SHPL=6 | Avg. SHPL=8 | Avg. SHPL=9 |
| 1 | 87.95 | 107.50 | 136.82 | 166.14 |
| 2 | 68.02 | 87.51 | 116.78 | 146.08 |
| 3 | 48.31 | 67.80 | 97.05 | 126.29 |
| 4 | 28.76 | 48.16 | 77.33 | 106.58 |
| 5 | 9.27 | 28.66 | 57.75 | 86.87 |
| 6 | 6.40 | 9.16 | 38.25 | 67.34 |
| 7 | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |

Another simulation was done for different combinations of Schedule Pressure Level and User-Practitioner Relationship level which indicates that Average Rework Cost may increase with increase in Average Schedule Pressure Level but this happens at an increased User-Practitioner Relationship level (Table 2).

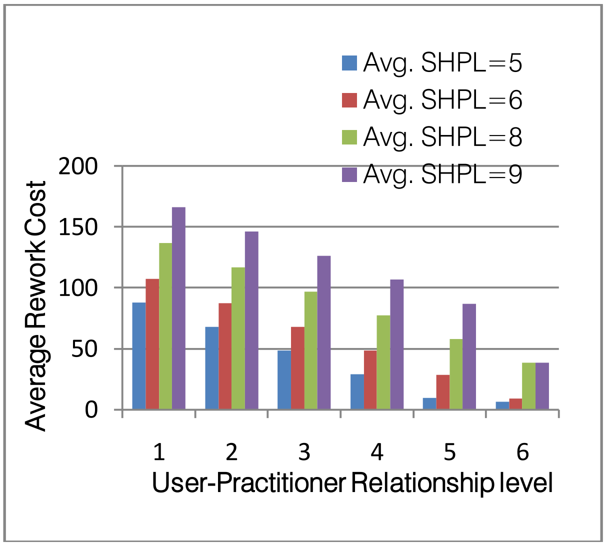


*Figure 3 :* Variation of Rework Cost with User-Practitioner Relationship Level at various schedule pressure levels

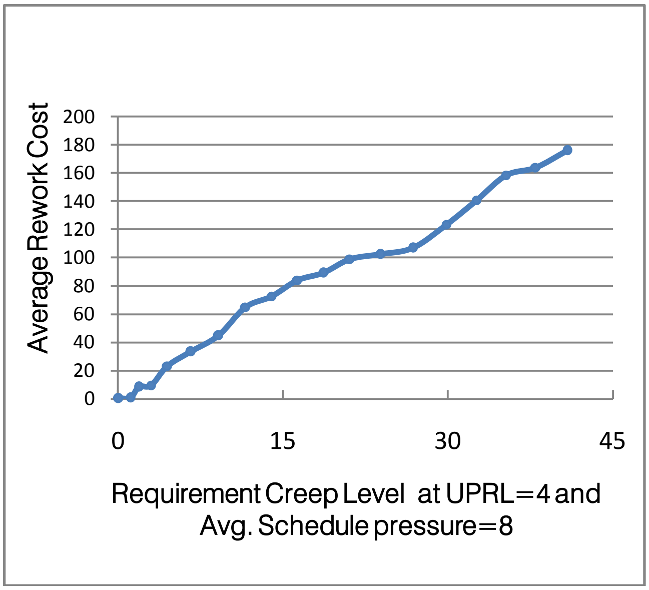


*Figure 4 :* Variation of Rework Cost with Requirements creep level at certain Schedule pressure level

An increase in requirements creep level may result into an increase in average rework cost. The increase becomes sharper at higher levels of requirements creep.

## VII. Conclusion

The stochastic simulator presented here in this paper models the potential effects of requirements creep as a risk factor on various software risk management factors. This will enable software project managers to take decisions in planning and scheduling the various activities involved in software development and perform sensitivity analysis in order to achieve the desired risk mitigation goals.

## References Références Referencias

1. Marc I. Kellner, Raymond J. Madachy, and David M. Raffo, "Software Process Modeling and Simulation: Why, What, How", Journal of Systems and Software, Vol. 46, No. 2/3, 1999.
2. Raymond J. Madachy, "A Software project dynamics model for process cost, Schedule and risk assessment", Ph.d. Dissertation, University of Southern California, 1994.
3. D. Houston, "A Software project simulation model for risk management", Ph.D. Dissertation, Arizona State University, Tempe, AZ, 2000.
4. B.W. Boehm, "Software risk management principles and practices", IEEE Software 8 (1), 32-41, 1991.
5. C. Jones, "Minimizing the risks of software development", Cutter IT Journal 11 (6), 13-21, 1998.
6. D. Gause and G. Weinberg, "Exploring Requirements", New York, NY, Dorset House, 1989.
7. D. Leffingwell and D. Widrig, "Managing Software Requirements - A Unified Approach", Addison Wesley Longman, Inc., 2000.
8. B. Curtis, H. Krasner and N. Iscoe, "A field study of the Software Design Process for Large systems", Communications of the ACM, 31(11), pp. 1268-1286,1988.
9. M. Lubars, C. Potts and C. Richter, "A review of the state of the practice in requirements modeling", Proc. of IEEE Symposium on Requirements Engineering, San Diego, Califonia,1993.
10. The Standish Group, The Standish Group Report - CHAOS, http://www.scs.carleton.ca/~beau/PM/Standish-Report.html, 2001-03-07, 1995.
11. K. R. Linberg, "Software developer perceptions about software project failure: a case study", The Journal of Systems and Software", 49(2-3):177–192, 1999.
12. J. Ropponen and K. Lyytinen, "Components of software development risk: How to address them? a project manager survey", IEEE Transactions on Software Engineering, 26(2):98–112, 2000.
13. R. Schmidt, K. Lyytinen, M. Keil, and P. Cule, "Identifying software project risks: An international delphi study", Journal of Management Information Systems, 17(4):5–36, 2001.
14. L.Wallace, M. Keil, and A. Rai, "Understanding software project risk: a cluster analysis", Information and Management, 42(1):115–125, 2004.
15. D. Houston, G. Mackulak and J. Collofello, "Stochastic simulation of risk factor potential effects for software development risk management", The Journal of Systems and Software 59 (2001) 247-257, 2001.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Cost based Model for Big Data Processing with Hadoop Architecture

By Mayank Bhushan & Sumit Kumar Yadav

*Motilal Nehru National Institute of Technology, India*

*Abstract-* With fast pace growth in technology, we are getting more options for making better and optimized systems. For handling huge amount of data, scalable resources are required. In order to move data for computation, measurable amount of time is taken by the systems. Here comes the technology of Hadoop, which works on distributed file system. In this, huge amount of data is stored in distributed manner for computation. Many racks save data in blocks with characteristic of fault tolerance, having at least three copies of a block. Map Reduce framework use to handle all computation and produce result. Jobtracker and Tasktracker work with MapReduce and processed current as well as historical data that's cost is calculated in this paper.

*Keywords: big data, hadoop, cloud computing, mapreduce.*

*GJCST-C Classification: H.2.8*

# Cost based Model for Big Data Processing with Hadoop Architecture

Mayank Bhushan [α] & Sumit Kumar Yadav [σ]

*Abstract-* With fast pace growth in technology, we are getting more options for making better and optimized systems. For handling huge amount of data, scalable resources are required. In order to move data for computation, measurable amount of time is taken by the systems. Here comes the technology of Hadoop, which works on distributed file system. In this, huge amount of data is stored in distributed manner for computation. Many racks save data in blocks with characteristic of fault tolerance, having at least three copies of a block. Map Reduce framework use to handle all computation and produce result. Jobtracker and Tasktracker work with MapReduce and processed current as well as historical data that's cost is calculated in this paper.

*Keywords: big data, hadoop, cloud computing, mapreduce.*

## I. Introduction

Technologies are changing rapidly, with lot of competition. In past, hardware cost was meaningful, as storage was a big issue for technological development, because of it's cost. Software and hardware, both having same cost at that time. After that software becomes complex in terms of development, but easy to use. Nowadays, with decrement in cost of hardware, the limitations of storage is not an issue. As functional programming, works with several functions [1] , so it requires large amount of space to run a program, reducing the execution time to a great extent[2]. So today's scenario is about faster execution without focusing on hardware cost. As industry is growing, hardware cost is getting lowered so sufficient amount of storage is available without difficulties. Earlier technologies were having specific views on hardware usage, now even 1TB is not a big deal for our commodity system.

Many social network use Resource Description Framework (RDF) [3]. Facebook's Open Graph [4], Freebase [5] and DBpedia [6] are having structured data. Facebook's Open Graph [4] show connection of user to its real functioning. Freebase [5] provide structured directories for music. DBpedia [6] provide structural contents from wikipedia. As per records till 2012, every minute usage of social networking site 'Facebook', having largest number of users, generating share of 684,478 pieces of contents, 'Youtube' users upload 48 hours video, 'Instagram' users share 3,600 new photos and 'Tumbir' sees 27,778 new post published [7]. A Boeing 737 engine generate 10 terabytes of data in every 30 minutes of flight [8]. All these data are information regarding weather conditions, positioning of plane, travelers information and other matters. So volume, velocity and complexity of data generation is increasing day to day. That require tool to handle it and more importantly with in time limit. Traditional database is not sufficient for doing all these calculation under the time limit. Here Hadoop fulfill all current requirements. Facebook, Google, LinkedIn, Twitter are establishing their business in Big Data. Many companies are still not having Hadoop professionals but they hire those from other companies. World's second largest populated country, India, having four times the population than USA, start trend of Big Data and is implementing Biometric System with unique ID number of every person. This project is called "Aadhar Project" that is world largest Biometric Identity project [9] with use of smart card technology and specification of International standard for electronic identification cards. With research perspective on Big Data, apart from Computer Science, other fields like Mathematics, Engineering, Business and Management, Physics and Astronomy, Social Science, Material Science, Medicine, Arts are also taking keen interest in that [10]. USA is on top, in research of Big Data issues, followed by China [10].

In today's world Big Data is moving towards cloud computing. Cloud computing provides required infrastructure as CPU, bandwidth, storage spaces at needed time. Organization like Facebook, LinkedIn, Twitter, Microsoft, Azure, Rackspace etc. have moved to cloud and doing Big Data analytic work, like Genome Project [11] that is processing petabytes of data in less amount of time. These technologies use MapReduce, for proper functioning. For moving Big Data to cloud, all data is moved and processed at data center [12], as being available at one place, cloud facilities can be easily provided.

In this paper section 2 is focusing on importance of MapReduce technique in current system and its practical uses there. Section 3 elaborate about features of Hadoop system with its functionalities.

*Author α: Department of Computer Science & Engineering, Motilal Nehru National Institute of Technology Allahabad, India.
e-mail: mayankbhushan20061@gmail.com*

*Author σ: Department of Computer Science & Engineering, Indira Gandhi Delhi Technical University; New Delhi, India.
e-mail: sumitarya0072@gmail.com*

Section 4 represent cost optimization while moving data to distributed environment. Section 5 concludes this paper.

## II. MapReduce: Visual Explanation

MapReduce is framework that work in distributed environment with server and client infrastructure. SPARQL is an RDF query language which used in social networking for data processing. SPARQL produce triplet as result of query process [3]. MapReduce provide functionality for processing query result. Facebook's close friend list, is output of processing of this technology in which 'selection' query processed then 'join' operation start functioning. Every 'join' process run one MapReduce function [13]. This is two layer mapping [3], refer to provide unnecessary MapReduce function for data processing. SPARQL generate triplet form of table in which 'selection' apply followed by 'join' operation. 'Selection' generate KEY-VALUE pair that is need for processing of MapReduce. Triplet ID is KEY assessment while its result is VALUE. Reduce function perform its functionality with same KEY-VALUE pair. 'Multiple join with filter' [3] proposed system with one layer mapping in which filter key used along with 'selection' and 'join' operations.

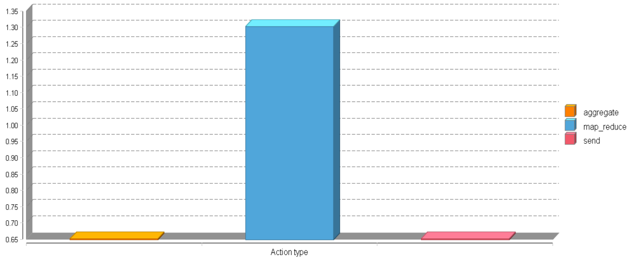


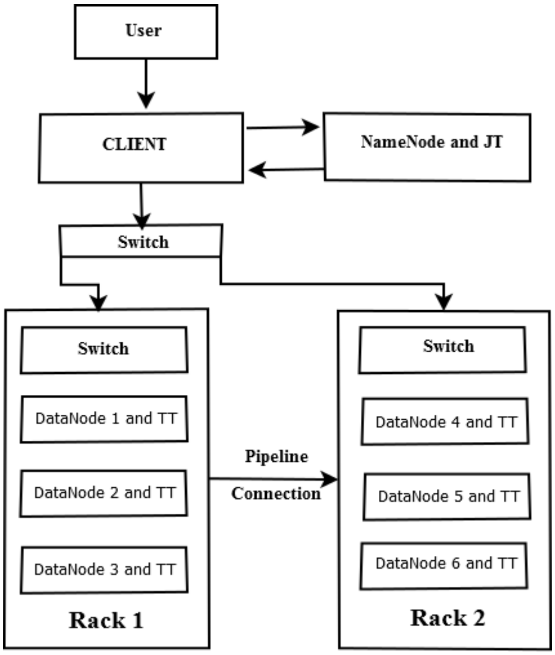
*Figure 1 :* MapReduce Analysis

Fig. 1 is showing analysis graph of MapReduce function with aggregation of data and sending the data using method of Map and Reduce. Taking an abstract model of Hadoop, MapReduce action is carried out with a rate of 1.65 per unit time, while aggregation and send actions are carried out at a rate of 0.65 per unit time.

MapReduce provides the services as text processing (wordcount, sort, terasort), web searching (pagerank) and machine learning (bayesian classification). HiBench [14] is providing MapReduce function to generate random data to include work load. MapReduce functioning consist four phases as 'map', 'shuffle', 'sort' and 'reduce'. 'Map' process generate intermediate result that need to be process further for resultant, 'reduce' phase start working preceded by shuffle and sort function. If there are 'P' no. of servers in cluster then shuffle phase has traffic O(P2) flows [15]. The standard concluding output size in Google jobs is 40.3% of the intermediate data set sizes. In the Facebook and Yahoo jobs consider in [16], the fall in size between the intermediate and the output data is even more distinct: in 81.7% of the Facebook jobs with a reduce segment, the final output data size is only 5.4% of the intermediate data size [15].

Server is responsible for assigning task for MapReduce. If there are 'P' systems and 'N' blocks of data then N/P blocks stored per system by server. Usually block size is user dependent and by default it is 64 MB. 'Map' phase generate (key, value) pair of data where each value have unique ID as key.

Server can run reduce function one time or more. It compute result based on (key, value) pair on server. Task, like web search query reduce function run one time that is sufficient for result [15].



*JT: JobTracker*
*TT: TaskTracker*

*Figure 2 :* Architecture of hadoop

They are several presented studies keen on the investigation of MapReduce procedure [17], [18], [19]. Yi Yuan et al. studied MapReduce with bases of CPU utilization, bandwidth, I/O of disk and network usage [20].

## III. Hadoop Framework System Model

In recent trends, Hadoop fixing its arms in software industry. Users of traditional database are keen to learn about it. Big Data use Hadoop framework for accessing the data. In 2012, IBM was biggest user of Big Data in revenues followed by HP, Teradata, Oracle, SAP, EMC, Amazon, Microsoft, Google, VMware, Cloudera, Hortonworks, Splunk, 10Gen and MapR [21]. Walmart leading the way with using Big Data on Hadoop for analyzing customer behavior and demand [22]. With huge amount of historical data as match records, individual records, conversations, meeting details etc., Australian open start using Big Data for analysis purpose; Netflix is largest commercial video provider in USA, start using Big Data on Hadoop [23]. Here discussion about architecture of Hadoop system with its key feature: Client, Master and slave node and HDFS.

### *a) Client*

Client is an application which used by end user and provide task to master and slave node for process. It ensure distributed data processing and distributed data storage. Apart from submitting job to cluster client machine it instruct for 'map' and 'reduce' and at last get the result as output. Client application accept job for processing and break it into blocks. Client take suggestion from master node about empty spaces and distributed these blocks to slaves.

### *b) Master Node and Slave Node*

Master node consist with Namenode and Jobtracker while slavenode consist with Datanode and Tasktracker as shown in fig. 2. Client ask Namenode about distribution of blocks. For safety of system block is replicated by minimum three. It is default replicas and it can be set further by user. Namenode provide list of Datanodes to client where data can be stored. Namenode stores meta data which store in RAM that consist information about all Datanodes, racks information, blank spaces, namespace of entire system like last modified time, create time, file size, owner permission, no. of replicas, block-ids and file name. Data retain in Datanode as it never fail; out of three copies one copy retain in by one Datanode in a rack whiletwo other copies put in another same rack but in different Datanodes. This feature gives the quality of fault tolerance with less chance of failure of Datanode and rack simultaneously. Transfer of all block is TCP connection so proper acknowledgment is there with pipeline processing with no wait for completion. Namenode keep updating its meta data as it receives acknowledgment from Datanode. Datanode keep sending signal with interval of three seconds indicating its aliveness; if it not receive by Namenode within 10 minutes then Datanode consider as dead and make it's replica to other node by master node.

If any file need to be executed then client ask Jobtracker to start executing file that reside in Hadoop Distributed File System (HDFS). Jobtracker takes information from Namenode about residence of operative blocks. After that Jobtracker instruct Tasktracker to run program for execution of file. Here 'map' function start and reported by signal to Jobtracker. Output of 'map' result store in Tasktracker's local memory. 'Map' results intermediate data and send it to a node which function by gathering all intermediate data for performing 'reduce' task. At last output is written to HDFS and sent to client.

### *c) Hadoop Distributed File System*

Hadoop use HDFS for storing the data that is distributed in nature and storing large data with streaming data pattern. Google file system (GFS) [24] also chunk based file system, use design of one master and many chunkservers. HDFS support fault tolerance with high throughput and can be built out of commodity hardware. But it is not useful for large amount of small files with low latency data access. GFS and HDFS do not execute POSIX semantics [25].

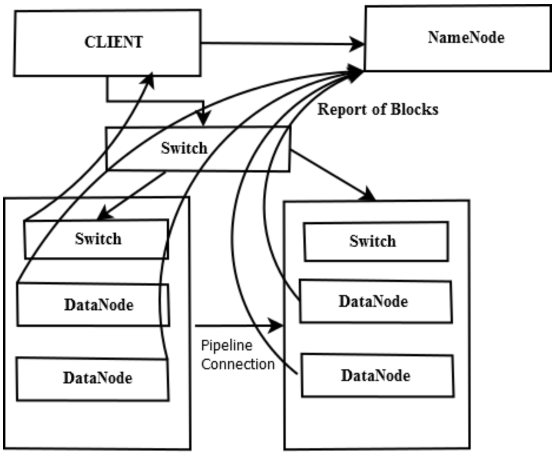


*Figure 3 :* Connection between Datanode and Namenode

## IV. Evaluation Cost in Hadoop Architecture

Consider a system where Client, Namenode, Datanode are connected. Let assume client (C) is connected to switches (P) in client side, Switches (Q) are in Datanode side where (D) numbers of Datanodes are connected to each other in a rack as fig. 3. These racks are connected as pipeline pattern. Such structure is reflected as architecture of Hadoop. Bandwidth between both switches is limited as $B_{P,Q}$

1. When any task comes to client for processing it consult with Namenode. Namenode regularly aware about rack storage for its availability with Datanodes. For engagement of further proceeding value XC;N take decision about connection signal between Namenode and client. Decision cost will be:

$$\text{Decision Cost}(X_{C,N}) = \begin{cases} 1 & \text{if X>0} \\ 0 & \text{if X=0} \end{cases} \qquad (1)$$

2. Client consult with Namenode which have information about rack system. Namenode having knowledge about which Datanode is free to occupy blocks of file which come to client for processing. This file is divided at least in three parts (up to user

choice). Namenode gives the address of maximum bandwidth rack first and continue with decreasing order of bandwidth. If assume data rate is _P;Q and total amount need to transfer is Gd(t) then bandwidth cost $B_{P,Q}$ will be:

$$\sum_{t=1}^{t=T}\eta_{P,Q}(\sum_{C,p\varepsilon P,q\varepsilon Q,d\varepsilon D} G_d(t)) \qquad (2)$$

Where p, q, d are one of the component from switches and Datanodes. This information store in RAM of Namenode. Gen2 Hadoop use secondary Namenode which access information for backup of Namenode's data from its RAM and store it to hard disk. Secondary Namenode is not a replacement of Namenode.

3. Datanode store information of current and historical data. As Datanode keep sending signal to Namenode about its aliveness through switch as fig. 3, if Datanode not send signal within 10 min to Namenode then Namenode assume it dead. Storage and estimation cost SSE will be:

$$\sum_{u=1}^{u=t-1}(\gamma)\eta_{P,Q}(u)+\sum_{u=t}^{u=T}\eta_{P,Q}(u) \qquad (3)$$

$(\gamma)$decide the factor of current or historical data. If any Datanode not sending signal from 10 min. then assume $(\gamma)$ to 0 but newly allocated data will be transferred to another node by estimation factor.

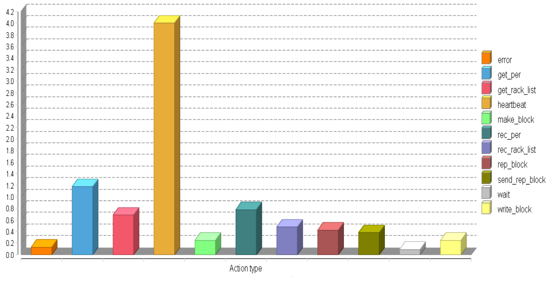


*Fig. 4 :* Performance Analysis

4. Jobtracker and Tasktracker that are associated with Namenode and Datanode respectively, do MapReduce function. Client load program of Map that executed by Jobtracker for finding situated targeted blocks after consulting by Namenode. Total distribution of blocks are less than number of racks. Tasktracker will produce result that might be 0. Now Reduce task will be executed which collect all intermediate result in a node. That node decide by Namenode and calculate result over there and transfer to HDFS. Resultant data Yred will be:

$$\sum_{u=1}^{u=T}R_{G_d}(u) \qquad (4)$$

5. Data move from client to Datanode after generated by user. This data will evaluate routing cost of data which included delay between clien to switches of user side, switches of user side to switches of Datanode side and switches of Datanode side to Datanode. Total routing cost $Z_{rct}$ will be:

$$\sum\zeta(G_d(t))(M_{C,P}+N_{P,Q}+O_{Q,D}) \qquad (5)$$

($M_{C;P}$ + $N_{P;Q}$ + $O_{Q;D}$) is pecuniary cost that showing latency from C ⇒ P ⇒ Q ⇒ D. $\zeta$ represent constant cost which convert weight cost to monetary cost. $(\zeta)$ depend on user as network use.

Fig. 4 is showing overall performance analysis on separate function with action type in X axis and normalized rate in Y axis. First is for error report of decision cost, second is about getting permission between Namenode and client. Third function showing rack list from Namenode. Forth is signal as heart beat which comes on master node in every three second, it is highest time response which happens frequently. Fifth pillar is showing work of client for dividing text file into blocks. Sixth pillar showing receive permission from Namenode, seventh is receiving rack list in which Datanode reside from Namenode so that chunks of file can be alloted. Eighth and Ninth showing replicas information. Data flow in racks as pipeline connection so least waiting rate as showing in tenth pillar. Last eleventh pillar showing action type for writing data into Datanode.

## V. Conclusion

This paper elaborated the architecture of Hadoop with its growing usage in industries as well as function of MapReduce on which current technologies moving. Among rack that consist of Datanodes and Tasktrackers choose by Namenode on basis of routing cost as showing in paper. It also evaluate cost of result that produce by different Datanodes. Decision of establishing communication of client with Datanode will also be decide by link between Namenode and client. Datanode may consist of historical data that's cost also get evaluated in this paper.

## References Références Referencias

1. Peter Henderson "Functional Programming Application and Implementation."
2. John Hughes "Why Functional Programming Matters." Institutionen for Datavetenskap, Chalmers Tekniska Hogskola.
3. Liu Liu, Jiangtao Yin, Lixin Gao "Efficient Social Network Data Query Processing on MapReduce." HotPlanet'13 Proceeding of the 5th ACM workshop on Hotplanet, Page(s): 2 - 32.
4. Open Graph, https://developers.facebook.Com/ docs/opengraph.
5. Freebase. http://www.freebase.com.
6. DBpedia. http://www.dbpedia.com.

7. www.visualnews.com/2012/06/19.
8. www.wipro.com/documents/big-data.pdf
9. www.dataname.com/dataname/2012-08-02
10. Human Genome Project, http://www.ornl.gov/hgmis/home.shtml.
11. www.researchtrends.com.
12. Linquan Zhang, Chuan Wu, Zongpeng Li, Chuanxiong Guo, Minghua Chen and Francis C.M.Lau "Moving Big Data to The Cloud.' Infocom, 2013 Proceeding IEEE, Turin, 14-19 April 2013, Page(s): 405 - 409.
13. P. Mika and G. Tummarello. "Web semantics in the clouds. Intelligent Systems." Intelligent System, IEEE, 23 Sep.2008, 23(5), Page(s)::8287.
14. S Huang, J. Huang, J. Dai, T. Xie and B. Huang. "The Hibench Benchmark Suite:Characterization of the MapReduce based Data Analysis." 26th International Conference on data Engineering Workshop (ICDEW), 2010, Page(s)::41-51.
15. P. Costa, A. Donnelly, A. Rowtron and G. O'Shea, "Camdoop: Exploiting in-network Aggregation for Big Data application." Proceeding USENIX NSDI 2012.
16. Chen, Y.Ganapathi, A.R.Griffith and Katz R., "The Case for Evaluating MapReduce Performance Using Workload Suite." Modeling, Analysis and Simulation of Computer and Telecommunication System, Singapore, 25-27 July 2011. Page(s)::390-399.
17. A. Pavlo, E. Paulson, A. Rasin, D.J. Abadi, D.J. Dewitt, S. Madden and M. Stonebraker. "A Comparison of Approaches to Large-Scale Data Analysis." Proceeding of the ACM SIGMOD International Conference on Management of Data, NewYork. Page(s):: 165-178.
18. M. Stonebraker, D. Abadi, D.J. Dewitt, S. Madden, E. Paulson, A. Pavlo and A. Rasin. "MapReduce and Parallel dbmss:Friends or Foes?" Communication of the ACM, January 2010. Page(s)::64-71.
19. D. Jiang, B.C. Ooi, L.Shi and S.Wu. "The Performance of MapReduce: An in-depth study." Proceeding of the VLDB Endowment. 1-2 September 2010. Page(s):: 472-483.
20. Yi Yuan, Haiyang Wang, Dan Wang, Jiangchuan Liu. "On Interference-aware Provisioning for Cloud-based Big Dat Processing." 21st International Symposium on Quality of Services. 3-4 June 2013. page(s)::1-6.
21. http://www.networkworld.com/slideshow/114134/.
22. http://www.gigaom.com/2011/07/17.
23. http://www.computerworld.com/slideshow.
24. S. Ghemawat, H. Gobioff and S. Leung. "The Google File System." SIGOPS Operating System Review. Page(s)::29-43.
25. Changquing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Kequie Li. "Big Data Processing in Cloud Computing Environments." 12th International Symposium on Pervasive System, Algorithm and Networks, San Marcos. 13-15 Dec. 2012. page (s) ::17-23.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Information Retrieval based on Content and Location Ontology for Search Engine (CLOSE)

By Niranjan Kumar & S. G. Raghavendra Prasad

*Rashtreeya Vidyalaya College of Engineering, India*

*Abstract-* This paper mainly focuses on the personalization of the search engine based on data mining technique, such that user preferences are taken into consideration. Clickthrough data is applied on the user profile to mine the user preferences in order to extract the features to know in which users are really interested. The basic idea behind the concept is to construct the content and location ontology's, where content represent the previous search records of the user and location refer to current location of user. SpyNB is the approach used to mining the user preferences from the Clickthrough data. The ranked support vector machine (RVSM) is performed on the searched results in order to display results according to user preferences by considering Clickthrough data.

*Keywords: SpyNB, personalization, ontology, RSVM, non-geographic search, geographic search, search engine optimization (SEO), personalized information retrieval (PIR).*

*GJCST-C Classification: H.3.3*

# Information Retrieval based on Content and Location Ontology for Search Engine (CLOSE)

Niranjan Kumar [α] & S. G. Raghavendra Prasad [σ]

*Abstract-* This paper mainly focuses on the personalization of the search engine based on data mining technique, such that user preferences are taken into consideration. Clickthrough data is applied on the user profile to mine the user preferences in order to extract the features to know in which users are really interested. The basic idea behind the concept is to construct the content and location ontology's, where content represent the previous search records of the user and location refer to current location of user. SpyNB is the approach used to mining the user preferences from the Clickthrough data. The ranked support vector machine (RVSM) is performed on the searched results in order to display results according to user preferences by considering Clickthrough data.

*Keywords: SpyNB, personalization, ontology, RSVM, non-geographic search, geographic search, search engine optimization (SEO), personalized information retrieval (PIR).*

## I. Introduction

In the modern information retrieval system, the results that are found should be more accurate to query submitted by the user, and also efficiency should be considered.

In order to solve the problems that are faced by the current search engine technology such as retrieving results that are irrelevant to the search query, the order in which they are displayed should be considered. According to Hele-Mai Haav [1] to solve problem of information retrieval in current information retrieval systems it should be improved by intelligence to manage the effective retrieval, filtering and presenting relevant information. So two main information retrieval models are classified as, keyword based information retrieval model and concept based information retrieval model. The indexing terms and Boolean logical queries are used in keyword based model, where indexing may be automatic or manual, when Boolean query are taken into consideration the frequency of occurrence is taken into account.

Context-aware system [2], depending on the user's relevancy the information/services is provided. For instance consider the keyword apple, it can mean as a fruit or it can mean as a mobile and laptops by Apple Company. When the query is submitted by two different users, irrespective of their interest same results are displayed for both users, if one user is interested only on apple accessories, for him both relevant and irrelevant information are displayed in random order. The information for what the user is looking may be in same document else somewhere in the overall document. The current system performs word to word matching of the search query.

Another instance in search engine is searching for places based on current location of the user. For example, if the user current location is Jaynagar and user trying to search restaurant near by current location, the search engine must show the restaurant which are near to the current location of the users and rest of the restaurant location other than jaynagar should be given next preference. The detailed discussion related to geographic and non-geographic search is given in proposed system section.

The main aspects that should be considered in information retrieval system is to reduce the complexity involved in query execution [3] such that performing lexical analysis, stemming process on the user query and construction of index terms. This paper focuses on search engine optimization (SEO) by reducing the complexity in the user query execution.

The rest of the paper is organized as: - In section II literature survey is carried out by surveying previous paper present, such that what are the technologies currently used to optimize the search engine. In section III technique to reduce the complexity for optimization of search query are studied. In section IV detailed view of implementation. In section V experimental evaluation and in IV Conclusion and future enhancements are discussed.

## II. Literature Survey

M. Rami Ghoran [4] studied that for every query that is submitted by the user he will get the relevant and irrelevant information for that query. So they classify the personalized information retrieval (PIR) system into three scopes: Individualized system, community-based system and aggregate-level system.

When individualized system is considered the system adaptive [5][6] decision are taken such that, the user interest and preferences are taken into account while

*Author α σ : RVCE, Bangalore. e-mails: niranjankumar213@gmail.com, raghavendrap@rvce.edu.in*

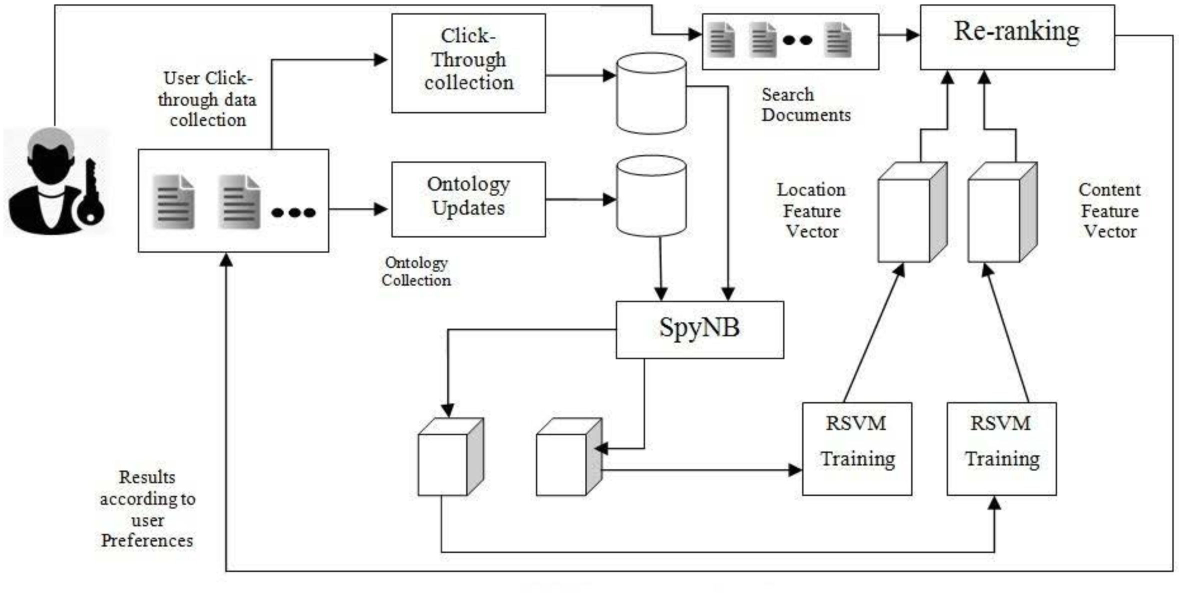


*Figure 1 :* Overall Architecture of the CLOSE system

Performing the search operations, while this approach leads to true to true personalization but it has some drawback such as:

Fresh start, when user is new to system his/her interest should be tracked and some time user may not compromise to share personal information with the system.

Community-based system [7] describes sharing of the information among several users/models. The data enrichment technique such as clustering technique is used in grouping of the similarity among various users. Using some similarity criteria the users among the web can be grouped into one model, so that results for this community can be personalized.

Aggregate-level system [8] where information gathered is represented in the form of summary for purpose of analysis. The common parameters such as age are considered to form clusters. For example a site selling music CD's may advertise certain CD's based on the age of the users and data aggregate for their age group. Online analytic processing (OLAP) is the simple type of data aggregation.

Browser also provides certain level of personalization by storing the cookies and recently visited web hyperlinks in the buffers. When the user is in static place browser will provide certain level of personalization, but when user place changes dynamically buffer contents are no more used.

For this purpose the new technique can be taken into consideration, such that each user's interest is maintained in the server buffer so that where ever user requests some result in form of query this can be compared with user interest buffer and relevant information can be retrieved from the system by minimizing unrelated results.

## III. System Design

Fig 1 shows the complete architecture of the CLOSE system, the working procedure is as follows. When the user is new to system and enters any query for the first time the preferences for location is taken along with search keyword and search operation is performed. The keyword of the query is searched in the server and relevant results are fetched and displayed as the results. When the user clicks on some links, Click through data will be recorded. Later when the user searches for the same keyword, the previously visited pages will be displayed first with higher ranked pages and, if there is are any new links they will be ranked in lower order.

Spy NB [9] is the algorithm used to fetch the user Click through data, and these are transformed to vectors for further process. The Ranked support Vector machine (RSVM) training is performed on the vectors for Re-ranking of search results according to user preferences. The detailed description about Spy NB and RSVM is given in implementation part.

The system mainly concentrates on building the method of ontology for all the possible keywords. The word can have different meaning in different context [2].

For example when the keyword "JAVA" is considered, in several perspectives it mean as the programming language, but by the name JAVA there is an island in Indonesia, and java coffee is referred to as a coffee beans.

When the two users submit the query both will get similar results either list of Java Island or list of java coffee beans is displayed or list of java programming is displayed, but one user expecting only about island and other only programming language. The system mainly focuses on differentiating which user is really interested in what. For this purpose the ontology is constructed for each keyword with their meaning. The fig 2 shows the construction of ontology for some words.

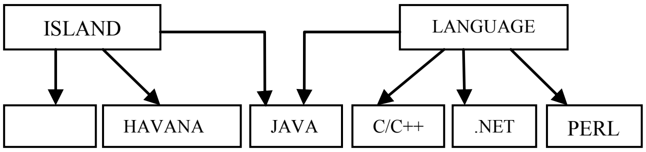


*Figure 2 :* Ontology for keyword JAVA

*Clickthrough data:* It is the process of recording the links or advertisement that is clicked by the user(s), for the purpose of determining which link is viewed how many times. The system makes use of these Clickthrough [10] data in personalizing each specific user's interest by maintaining the records for each user in the database. In formal language it can be defined as, it is triplets of (Q, R, C) where Q is the query, R is the ranking order in which it is displayed and C is the set of URLs that are clicked by the users. To achieve personalization the system is classified into two distinct levels namely, content ontology and location ontology [11] [12]. The detailed descriptions about two levels are elaborated in below section:

*Content Ontology:* The concept works on extracting the keywords/phrase from the web snippets by eliminating all the stems in the query Q. The content ontology is classified differently to different users based on their interest. The co-existence of the keyword in the query Q is calculated to find similarity among the user interest by using following support and confidence rule [3]:

$$Support(c_i) = \frac{sf(c_i)}{n}.|c_i|$$

Where $s\ f\ (c_i)$ is the web snippet frequency of the keyword/phrase in the query Q, n is the total number of web snippet and $|c_i|$ is the number of terms in the keyword/phrase $c_i$. If the support of the keyword/phase ci is higher than threshold ΔT (where threshold ΔT is set by user), than we consider $c_i$ as the concept for query Q.

In this system the value of ΔT is set to 5 because, if ΔT value is assigned with lesser value than for each search, ranking should be updated this leads to consume more time for reordering of links. If ΔT value is assigned with larger value than perfect personalization cannot be achieved.

The following two prepositions are adopted to find relationship between concepts for ontology:

- Similarity: The two concepts which coexist more in the search results can be considered or represented as the same topic of interest. If occurrence of document $c_i$, $c_j$ > ΔT (where ΔT is the threshold) then $c_i$ and $c_j$ can be considered as similar.
- Parent-Child Relationship: specific concepts appear with general terms, but backtracking is not true. If the preference of $c_i$ and $c_j$ > ΔT then we can conclude that $c_i$ is child of $c_j$.

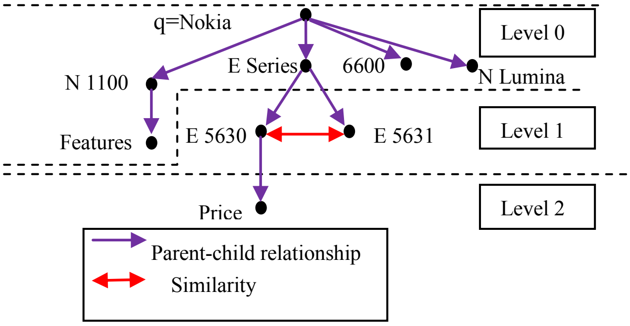


*Figure 3 :* Ontology's classification for q=Nokia

Fig 3 shows the content ontology for the query q=Nokia, where the concept linked with single head arrow indicates parent child relationship and double head arrow indicates similarity concepts. In the fig 2 the possible concept space determined for the keyword/phrase "Nokia" while Click through data will determine the preferences on the concept based. The concept space for the query "nokia" consists of different types of models such as E-series, N-Lumina etc. When E-series is taken into consideration both has similarity that they belong to same parent.

Content space for the query "Nokia" consists of "N1100", "E-series", "6600", and so on. If the user is interested in E-series and clicks on the page containing price, the Click through of the links are captured. These Click through data is considered as the positive preferences and vector is constructed.

When the same query is issued by the same user later the vector is transferred to server by transforming this content vector into content weight vector to rank the search result according to user preferences.

*Location Ontology:* The approach of the location ontology [13] [14] [15] is quite different from the construction of content ontology. Following assumptions are made i.e., the parent-child relationship cannot be accurately derived for the location ontology. To construct the vector [15] for location concept following Bangalore, "Jaynagar/Bangalore/Karnataka/India", is associated with the document d.

The construction of the vector for the location ontology is similar to that of the content ontology. The Clickthrough data is transferred to the server and transformed as the location vector and this vector is used to rank the user preferences.

## IV. Implementation

In this section technique that are used to personalize the search engine are discussed in detail. First, when the query q is entered by the user, look for previous records if previous search results are found then apply Content ontology concept else if the user is new then accept the query q and apply Location ontology concept.

**Algorithm 1**: CLOSE (Ui, q, L)

```
// Input: User identity Ui, Query q and Current location of User L.
// Output: Results for query with user preferences.
1. Accept the Query q from user where q ϵ {A-Z, a-z, 0-9}
2. Filter the post (documents) using the keyword q
If (∀ Post (di) == compare (q))
3. If (check user profile Ui for previous records)
4. Result_set ⟵ Content-Ontology (Ui, q) + Location-Ontology (Ui, q, L)
5. Update Ui ⟵ Result_set
Display “Results”
6. Else
7. Result_set ⟵ Location-Ontology (Ui, q, L)
8. Update Ui ⟵ Result_set
Display “Results”
9. End if
10. Else
11. Display “Query Not Matched”.
```

Next algorithm will be related to searching keyword based on Content ontology.

**Algorithm 2**: Content-Ontology (Ui, q)

```
// Input: User Identitȳ, and corresponding Query q.
// Output: Return Results to CLOSE
1. Let S ⟵ post (di) matched for q.
2. Retrieved ⟵ SpyNB(S).
3. Let Ps denotes Positive set and Ns denotes Negative set from SpyNB(S) where:
Ps ϵ {Links that are clicked by the users}
Ns ϵ {Links not clicked by the users}
Select Positive Set from Retrieved documents.
4. Count ⟵ Count+ Number_of_clicks.
5. Results ⟵ RSVM (Count, post_code).
6. Return Results.
```

Next algorithm will be related to searching keyword based on Location ontology.

**Algorithm 3**: Location-Ontology (Ui, q, L)

```
// Input: User-Identity $U_i$, Query q and Location L.
// Output: Return Results to CLOSE.
1. Let L ⟵ Current Location of User.
L1      Post-Location.
2. Let S ⟵ Post (di) matched with q && L
3. Calculate distance between current location and Post Location
Difference ⟵ L-L1
4. Result ⟵ Sort post with shortest distance to Higher Distance.
5. Return Result
```

Spy Naive Bayes (SpyNB) algorithm is used to collect the Clickthrough data. This algorithm will maintain two sets called positive set Ps and negative set Ns. Where

$P_s$ ϵ {Links that are clicked by the users}
$N_s$ ϵ {Links not clicked by the users}

**Algorithm 4**: SpyNB(s)

```
// Input: Post matched for Query q.
// Output: Feature vector for Post
1. Compare S with the user record.
2. If (S ϵ Ui)
3. Select post from the records.
Relevant_Post ⟵ Post (d  i).
4. Construct the Positive set and Negative set
5. Update Positive set in corresponding User Buffer.
6. Repeat for all Query q
7. End if
8. Return Post
```

Ranking algorithm will rank the results according to the user preferences by calculating the weight of both content and location concepts, for keyword/ key phrase. The content weight of all posts for particular keyword is considered in calculating the ranking order.

The vector support machine is constructed for training the user preferences, loop is entered when the ranking operation is started, and the number of count is recorded for the link whenever the user clicks on it. When the post reaches the minimum threshold value then it will gain a higher order value as compared from rest of the post. The formal representation for performing these is depicted below:

**Algorithm 5**: RSVM (count, post_code)

```
// Input: count for each click is taken as the input.
// Output: Ranking order of the posts.
1. For i ⟵ 0 to total_post-1 do
2. Content_weight_count ⟵ count.
3. Calculate the Content weight for particular keyword.
   P_code ⟵ Post_code
4. Content_weight (%) ⟵ $\frac{Pcode\_content\ weight\ Count}{\sum_{i=1}^{n} Conent\ weight\ count}$
5. Final_content_weight ⟵ $\frac{Content\_Weight}{2}$
6. P1 ⟵ $(location\ )/\ \sum_{i=1}^{n} total\ distance$
7. P2 ⟵ P1-100
8. location_weight_parameter ⟵ $\frac{P1+P2}{2}$
9. Final_rank ⟵ Final_content_weight
   + location_weight_parameter
```

## V. Experimental Evaluation

The Table 1 gives the dataset of the content ontology construction for some of the keywords. The table mainly consists of unique code for particular root keyword, name of keyword and parent of the corresponding keyword [17].

*Table 1 :* Statistic of Content Ontology

| Unique Code | Keywords | Parent |
|---|---|---|
| 101 | Hotel | 0 |
| 102 | Reservation | 101 |
| 103 | Facilities | 101 |
| 104 | Meeting Room | 103 |
| 105 | Party Hall | 103 |
| 106 | Animal | 0 |
| 107 | Jaguar | 106 |
| 108 | Lion | 106 |
| 109 | Car | 0 |
| 110 | Jaguar | 109 |
| 111 | BMW | 109 |
| 112 | Black Jaguar | 107 |
| 113 | Elephant | 106 |

Unique Code Keywords Parent 101 Hotel 0 102 Reservation 101 103 Facilities 101 104 Meeting Room 103 105 Party Hall 103 106 Animal 0 107 Jaguar 106 108 Lion 106 109 Car 0 110 Jaguar 109 111 BMW 109 112 Black Jaguar 107 113 Elephant 106

In the experimental evaluation "Hotel" is the root word and it has four children such as "Reservation", "Facilities", "Meeting Room", and "Party hall", similarly for others also constructed.

Similarly Table 2 gives the dataset of the location ontology construction for some of the locations. The table mainly consists of location code, Location name, latitude, longitude and parent of location. When location is considered, boundary value of 11 values is taken into consideration.

*Table 2 :* Statistic of Location Ontology

| Location Code | Location Name | Parent | Latitude | Longitude |
|---|---|---|---|---|
| 1 | India | 0 | 21.0 | 78.0 |
| 12 | Karnataka | 200 | 12.97 | 77.56 |
| 123 | Bangalore | 201 | 12.97 | 77.57 |
| 124 | Mysore | 201 | 12.303106 | 76.640228 |
| 1231 | Jaynagar | 202 | 12.93 | 77.6 |
| 1232 | Koramangala | 202 | 12.933881 | 77.622343 |
| 13 | Tamil Nadu | 200 | 13.08 | 80.27 |
| 2 | London | 0 | 51.51 | -0.12 |
| 21 | Barking and Dagenham | 207 | 51.545268 | 0.147575 |
| 22 | Barnet | 207 | 51.650194 | -0.200897 |
| 23 | Bexley | 207 | 51.441811 | 0.154297 |

In posting of documents the related information are stored by entering the root and location for which it belongs. In this case Hotel "comfort" comes under Bangalore city for which India will be root, and so on others are posted.

When user enters the query q, the searching process will be carried out as mentioned in the implementation section by invoking several techniques. When the corresponding documents are found, and previous records of users are analyzed, the ranking support vector machine is performed on the posts that are matched by the keyword or query q.

Table 3 gives the RSVM calculation for the Keyword "jaguar for two different users, it can be observe from the table that two user have their own preferences in choosing the link.

Later, when two users search for same keyword then threshold value changes and ranking of their search results will be altered.

*Table 3 :* RSVM training of the Data sets

| Keyword | Posting number | Count | content Weight | Final Content Weight | Location | Distance | $P_1$ | $P_2$ | Final Location Weight | Final Value | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Jaguar User1 | 1001 | 0 | 0 | 0 | Jaynagar | 0 | 0 | 100 | 50 | 50 | 3 |
| | 1002 | 10 | 58.82 | 29.41 | Mysore | 160 | 18.47 | 81.52 | 40.76 | 70.17 | 1 |
| | 1003 | 5 | 29.41 | 14.70 | Koramangala | 6 | 0.69 | 99.30 | 49.65 | 64.35 | 2 |
| | 1004 | 2 | 11.76 | 5.88 | Delhi | 700 | 80.83 | 19.16 | 9.58 | 15.46 | 4 |
| | | Total=17 | | | | Total=866 | | | | | |
| Jaguar User2 | 1001 | 0 | 0 | 0 | Mysore | 0 | 0 | 100 | 50 | 50 | 3 |
| | 1002 | 5 | 29.41 | 14.70 | Jaynagar | 160 | 18.47 | 81.52 | 40.76 | 55.46 | 2 |
| | 1003 | 5 | 29.41 | 17.70 | Koramangala | 6 | 0.69 | 99.30 | 49.65 | 64.35 | 1 |
| | 1004 | 1 | 5.88 | 2.94 | Delhi | 700 | 80.83 | 19.16 | 9.58 | 12.52 | 4 |
| | | Total=11 | | | | Total=866 | | | | | |

## VI. Conclusion and Future Enhancement

We can conclude that the CLOSE system will provide better search results as compared to rest of the search engines by considering the users Content and location concepts. CLOSE system will take user preferences in minimizing the possible time for retrieving search results. RSVM training will be performed for each individual user profile, so that system will come to know in what the user is really interested.

As a future enhancement it can be extended by considering time as one of the parameter to even more optimize the search results. The sessions can also be considered as one of the parameter, so that when user stop work at particular instance, later when user get into system, at moment where user stopped working or viewing content of some documents, from that session it should be started (with respect to two or more different systems).

## VII. Acknowledgement

Foremost, I would like to express my sincere gratitude to my guide Mr. S G Raghavendra Prasad Assistant Professor, ISE Dept, RVCE, for the continuous support of my M. Tech study, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of writing this technical paper.

Besides my guide, I would like to thank the rest of my M.Tech committee: Dr. Jitendranath Mungara PG Dean, ISE Dept, RVCE, and Dr. Cauvery N K. HOD ISE Dept, RVCE.

## References Références Referencias

1. Hele-Mai Haav, Tanel-Lauri Lubi "A Survey of Concept based Information Retrieval Tools on the Web" White paper.
2. Deepika Bhatia et al "Context-aware Personalized Mobile Web Search Techniques-A Review" vol. 2(5), (IJCSIT) International Journal of Computer Science and Information Technologies, 2011, pp. 2440-2443.
3. Baeza-Yates R., RIBEIRO-NETO, B. 2013 Modern Information Retrieval: The Concepts and Technology behind Search, Pearson Edition.
4. M. Rami Ghorab et al "Personalised information retrieval: survey and classification", Centre for Next Generation Localisation Knowledge & Data Engineering Group. Pp. 1-40.
5. Kanika Arora, Kamal kant "Techniques for Adaptive websites and Web Personalization without any user effort" IEEE Students conference on Electrical, Electronics and Computer Science, 2012.
6. Athanasios Papagelis, Christos Zaroliagis "A Collaborative Decentralized Approach to Web Search" vol. 42 No. 5, IEEE Transaction on Systems, Man, and Cybernetics , 2012, pp. 1271-1290.
7. Dou Shen et al "Query Enrichment for Web-query Classification" vol. 24 No. 3, ACM Transactions on Information Systems, 2006, pp. 320-352.
8. Bamshad Mobasher et al "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization" Vol. 6 No. 1, ACM Transaction on Data Mining and Knowledge Discovery, 2002, pp. 61-82.
9. Wilfred Ng et al "Mining User Preference Using Spy Voting for Search Engine Personalization" Vol. 7 No. 3, ACM Transaction on Internet Technologies,2007, pp. 1-28.
10. Venington .K, R. Shanmugalakshmi "Enhancing personalized web search re-ranking algorithm by incorporating user profile" IEEE (ICCCNT), 2012, pp. 1-6.
11. LI Qing-shan et al "Ontology based User Personalization Mechanism in Meta Search Engine" IEEE (URKE), 2012, pp. 230-234.

12. Abdelkrim Bouramoul et al "An ontology-based approach for semantics ranking of the web search engines results" IEEE (ICMCS), 2012, pp. 797-802.
13. Varun Mishra et al "Improving Mobile Search through Location Based Context and Personalization" IEEE (ICCSNT), 2012, pp.392-396.
14. Sandeep Jain, Aakanksha Mahajan "Data Mining Based on Semantic Similarity to Mine New Association Rules" Vol. 12 Issue 12 Version 1.2 Global Journal of Computer Science and Technology Software & Data Engineering, 2012.
15. Mingyang Sun et al "FoSSicker: A Personalized Search Engine by Location-Awareness" IEEE (ICNC), 2012, pp. 456-460.
16. Al Sharji Safiya et al "Enhancing the Degree of Personalization through Vector Space Model and Profile Ontology" IEEE Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF, 2013, pp. 248-252.

    Shikha Goel et al "Search Engine Evaluation Based on Page Level Keywords" IEEE (IACC), 2013, pp. 870-876.
17. Vishwas Raval, Padam Kumar "SEReleC (Search Engine Result Refinement and Classification) - A Meta Search Engine based on Combinatorial Search and Search Keyword based Link Classification" IEEE (ICAESM), 2012, pp. 671-631.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Mining Health Care Sequences using Weighted Associative Classifier

By Sunita Soni & Dr. O. P. Vyas

*Bhilai Institute of Technology, India*

*Abstract-* This paper proposes the general framework for mining sequences from health care database. The database is a relational model consisting of set of temporal records of individual patient consisting of basic information of the patient ie Patient_ID, age, gender etc. the second part is a series of sequences representing the set of treatment given to the patient during regular visit to the doctor and the third part is class label. Similarity search of sequences is performed to convert the database of sequences, to the database of items, so that apriori algorithm can be applied. Weighted association rule mining has been performed to find the frequent sequence of treatment provided to the patient. Classification association rules (CAR) having positive class label as consequent, represents the frequent sequence of treatment given to the patient for successful treatment. With the experimental results, author feels confident in declaring that the framework is feasible in the medical domain.

*Keywords: sequence mining, weighted associative classifier, weighted support, weighted confidence, prediction*

*GJCST-C Classification: H.2.8*

MININGHEALTHCARESEQUENCESUSINGWEIGHTEDASSOCIATIVECLASSIFIER

*Strictly as per the compliance and regulations of:*

# Mining Health Care Sequences using Weighted Associative Classifier

Sunita Soni [α] & Dr. O. P. Vyas [σ]

*Abstract-* This paper proposes the general framework for mining sequences from health care database. The database is a relational model consisting of set of temporal records of individual patient consisting of basic information of the patient ie Patient_ID, age, gender etc. the second part is a series of sequences representing the set of treatment given to the patient during regular visit to the doctor and the third part is class label. Similarity search of sequences is performed to convert the database of sequences, to the database of items, so that apriori algorithm can be applied. Weighted association rule mining has been performed to find the frequent sequence of treatment provided to the patient. Classification association rules (CAR) having positive class label as consequent, represents the frequent sequence of treatment given to the patient for successful treatment. With the experimental results, author feels confident in declaring that the framework is feasible in the medical domain.

*Keywords: sequence mining, weighted associative classifier, weighted support, weighted confidence, prediction.*

## I. Introduction

Time plays a crucial role as patient's care as well as data collection and decision-making activities are performed over time. It is therefore often mandatory to deal with the temporal aspects by deriving useful summaries of the patient's behavior, including physiological signals or measurement time series, and adapting the decisions to the accumulated data and information. The goal of predictive data mining is to derive models that can use patient's historical information to exploit hidden information which will ultimately improve clinical Decision-making [1].

Diagnosis is related to the classification of patients into disease classes or subclasses on the basis of patients' data gathered from regular visit gathered time series. There are a growing number of papers that exploit data mining approaches for clinical prediction purposes. In a clinical context, predictions may support diagnostic, therapeutic, or monitoring tasks. Therapeutic prediction means the choice of the most suitable treatment for the patient.

Time series or temporal sequences; appear naturally in a variety of different domains, from engineering to scientific research, finance and medicine. In healthcare, temporal sequences are a reality for decades; with data originated by complex data acquisition systems like ECG's or even with simple ones like measuring the patient temperature or treatments effectiveness . In the last years, with the development of medical informatics, the amount of data has increased considerably, and more than ever, the need to react in real-time to any change in the patient behavior is crucial. In general, applications that deal with temporal sequences serve mainly to support diagnosis and to predict future behaviors [2].

The ultimate goal of temporal data mining is to discover hidden relations between sequences and subsequences of events. Just to mention few, following prediction can be performed using patient's historical temporal data.

1. Prediction for drug treatment planning or for the prognosis of surgical interventions.
2. Predictions in clinical monitoring are crucial in several contexts, such as in intensive care units (ICUs), which needs continue updating on the basis of the monitoring data.
3. Prediction may range from simply predicting the risk of disease based on the age factor or lifestyle for whole population, to the forecast of consequences of taking a particular drug or treatment for long time. For example drug taken for hypertension for a long time may affect the functioning of kidney.
4. Prediction of chance of any disease or casualty on neonatal based on different symptoms and other information like weight, systematic growth mother's blood group etc.
5. Predicting the risk of chronic disease as a result of another disease. For example diabetic patient having hypertension are more prone to Cardio Vascular Disease.

In this paper we have proposed a general framework to mine prediction rule for the accurate treatment of disease which will ultimately lead to cure of disease. The framework uses the historical data of the patient consisting of sequence of treatment given at regular interval. Further each sequence element may be various pathological test or advanced test results, regular observations like blood sugar, blood pressure etc. and medicines and other treatment recommended to the patient for that time period. The database consists of set of sequence of treatment given to the patient and a class label that defines whether the patient is cured or not.

*Author α : Department of computer applications, Bhilai Institute of Technology. e-mail: sunitasoni74@gmail.com*
*Author σ : Professor, Indian Institute of Information Technology, Allahabad, India.*

The major steps of the work proposed are

1. *Representation and modeling:* In this step, sequences of the temporal data are transformed into a suitable form. Every unique sequence is assigned a numeric symbol using step 2 and ultimately the database is converted to form suitable to perform apriori type algorithm.
2. *Similarity Measure:* This step defines the similarity measures between sequences. We are using Euclidian distance measure to find the unique sequences.
3. *Mining Operation:* In this step actual mining operation is performed to extract hidden information. In this framework we are extracting the set of frequent sequences (representing the treatment given to the patient) applied on the patient, which ultimately caused the patient to be cured. Association rule mining is used to find the association among the treatment given, with the given class label, the rules in this step are known as Class Association Rule (CAR).
4. *Prediction:* We use the high confidence CAR rules generated in step 3 to predict the sequence of treatment.

The proposed Framework for sequence mining is shown in figure1.

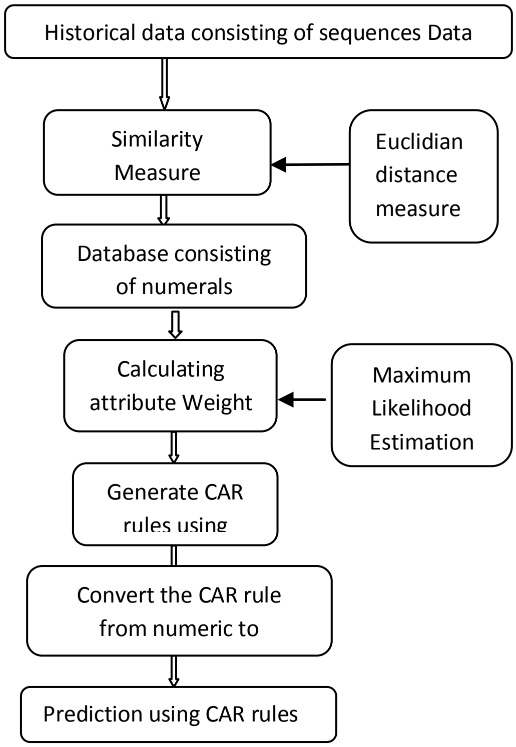

Historical data consisting of sequences Data
Similarity Measure
Euclidian distance measure
Database consisting of numerals
Calculating attribute Weight
Maximum Likelihood Estimation
Generate CAR rules using
Convert the CAR rule from numeric to
Prediction using CAR rules

*Fig 1 :* Sequence Mining framework using WAC.

## II. Related Work

### *a) Medical Prediction Using Temporal Data Mining*

Temporal databases consist in databases containing time stamped information. A time stamp can be represented by a valid time, which denotes the time period in which the element information is true in the modeled real world, and/or by a transaction time, which is the time in which the information is stored in the database. Temporal data mining approaches provide the opportunity to address different tasks, such as data exploration, clustering, classification or prediction [3].

In [4] temporal data mining has been applied on the hepatitis temporal database collected at Chiba university hospital between 1982-2001. The database is large where each patient record consists of 983 tests represented as sequences of irregular timestamp points with different lengths. The work presents a temporal abstraction approach to mine knowledge from this hepatitis database.

In [5] visual data mining technique on temporal data is applied for the management of hemodialysis. The approach is based on the integration of 3D and 2D information visualization techniques which offers a set of interactive functionalities. The paper described the main features of IPBC (Interactive Parallel Bar Charts), a VDM system developed to interactively analyze collections of time-series, and showed its application to the real clinical context of hemodialysis.

In [6] temporal data mining techniques has been applied to extract information from temporal health records consisting of a time series of elderly diabetic patients' tests. The first step is to find pattern structures using structural-based search using wavelets. In the second step a value-based search over the discovered patterns using the statistical distribution of data values is performed. In the third step the results from the first two steps is combined to form a hybrid model. The feature of the hybrid model proposed is the expressive power of both wavelet analysis and the statistical distribution of the values. Global patterns have been identified successfully.

In [7] initially a framework is proposed for the definition of methods and tools for the assessment of the clinical performance of hemodialysis (HD) services, on the basis of the time series which has been automatically collected during hemodialysis sessions. For the implementation the method proposed is intelligent data analysis and temporal data mining techniques to gain insight and to discover knowledge on the causes of unsatisfactory clinical results.

In [8] a new kind of temporal association rule and the related extraction algorithm is proposed. An Apriori-like algorithm has been used to search for meaningful temporal relationships among the complex patterns of interest.

In [9] a new algorithm is presented to mining of Temporal Association Rules which has the main innovative feature of handling both events with a temporal duration and events represented by single time points. This new method has been applied to analyze the healthcare administrative data of diabetic patients.

The method is found to be useful to observe frequent health care temporal patterns in a population.

In [10] a general methodology for the mining of Temporal Association Rules on sequences of hybrid events is proposed. The experimental results show that the method can be a practically used for the evaluation of the care delivery flow for specific pathologies. In [11] the work done in[10] has been extended to focus on the care delivery flow of Diabetes Mellitus, and an algorithm is proposed for the extraction of temporal association rules on sequences of hybrid events. This work has been extended in [12] to show how the method can be used to highlight cases and conditions which lead to the highest pharmaceutical costs. Considering the perspective of a regional healthcare agency, this method could be properly exploited to assess the overall standards and quality of care, while lowering costs.

In [13] an efficient technique to match and retrieve the sequence of different lengths has been proposed. A number of the research works proposed earlier were concentrated on similarity matching and retrieval of sequences of the same length using Euclidean distance metric. In the matching process a mapping among non-matching elements is created to check for the unacceptable deviations among them. An indexing scheme is proposed for efficient retrieval of matching sequences, which is totally based on lengths and relative distances between sequences.

In [14] the analysis of sequential data streams to unearth any hidden regularities is discussed and also the applications of it in various field ranging from finance to manufacturing processes to bioinformatics is explained. The notions of sequential patterns or frequent episodes represent only the currently popular structures for patterns. The field of temporal data mining is relatively young hence new developments in the near future is yet to come. The paper discuss such several issues and others like what constitutes an interesting pattern in data, problem of defining data structures for interesting patterns, linking pattern discovery methods etc.

### *b) Association Rule Based Classifier*

Given a set of cases with class labels as a training set, classification is to build a model (called classifier) to predict class label of future data objects. Associative classification is an integrated framework of association rule mining and classification. A special subset of association rules whose right-hand-side is restricted to the classification class attribute is used for classification. This subset of rules is referred as the class association rules (CAR). Extensive performance studies show that association based classification may have better accuracy in general [15], [16], [17]. The major advantages of new Predictive Model over the other models are-

- Fast training mechanism regardless of the size of the training set.
- Training sets with high dimensionality can be handled easy.
- Classification can be very fast with a compact set of rules.
- The classification model is easily understandable to humans (interpretability) well-organized, and easier to use model.
- Provides better accuracy than traditional decision tree classification algorithms.
- In medicine we are interested in creating understandable to human descriptions of medical concepts, or models. Associative classifiers are used for achieving this goal, since they can create a model in terms of intuitively transparent rule of the form *X→Y*. On the other hand, unintuitive black box methods, like artificial neural networks, may be of less interest.

In section III we have discussed some basic definition for sequence mining. In section IV the different steps of sequence mining is discussed. In section V the algorithm weighted associative classifiers is discussed. In section VI conclusion and future work has been discussed.

## III. Problem Definition

*Definition1:* Sequence Database: A sequence database D is a set of records D[0], D[1],...,D[n] where record D[i] represents the record of ith patient consists of ordered sequences, S(i,1), S(i,2), S(i,3), …S(i,j),…, where each sequence S(i,,j) is observed at time stamp tj , 1≤ j ≤ n, n is positive integer. Si represents a sequence observed at time stamp ti. In database D the size of record may be varying because the number of visits for the complete treatment of one patient may be different from other patient.

*Example:* For patient 3 the number of sequence is i, whereas the number of sequence for patient 1, 2 and 4 is m.

*Definition 2:* Sequence: An ordered sequence si is set of elements ek, where 1≤ k≤,l,

i.e. Si = (e1, e2, e3…el)

Each element ek belongs to some domain representing quantitative or categorical or binary value corresponding to any preliminary test results like blood pressure, blood sugar, body mass index, or other pathological test results or medication recommendation based on the test result at time stamp $t_i$.

*Definition 3:* Sequence length: Length of sequence is defined as number of elements in the sequence.
length ($S_i$) = number of elements in $S_i$ .

*Definition 4:* Sequence Structure: Structure of sequence is defined as the length of sequence and the elements and their order in the sequence. The exact sequence length and structure of sequence will be based on the disease for which the training data is collected. A typical example of structure of sequence and sequence in case of heart patient may be-

*Example:* Structure of the sequence is (Blood_pressure_upper, Blood_pressure_lower, Fasting_Blood_Sugar, BMI, test1, test2, Medicine1, Medicine2, Medicine3) and corresponding sequence is (190,50, 150, result_test1, result_test2, med1, med2, med3).

*Definition 5:* The sequence for one patient at different time stamp may be same or varying, also the sequence at same time stamp for the different patient may be same or varying. i.e.

1. Let Si is a sequence at ti and Sj is sequence at tj and $S_i$, $S_j \in D[i]$ then $S_i = S_j$ is possible. Let at time stamp ti , Si ∈ D[0] and Sj ∈ D[1]

then Si≠Sj or Si= Sj is possible.

The operator = and ≠ are discussed in Definition 6.

2. *Example:* patient 2 and 4 have given same treatment at same time stamp, also patient 2 has been given same treatment from time stamp $t_1$ to $t_i$.

*Table 1 :* Relational database D with set of temporal records

| | Time Dimension→ | | | $t_1$ | … | $t_i$ | … | $t_m$ | |
|---|---|---|---|---|---|---|---|---|---|
| | P_Id | age | gender | $S_1$ | … | $S_i$ | … | $S_m$ | class_label |
| Patient Record→ | 1 | 45 | f | (190,50,150,result_test1, result_test2, med1) | … | (200,90, 150, result_test1, result_test2, med1, med2) | … | (200,90,150, result_test1, result_test2, med1, med2 | Disease_cured |
| | 2 | 30 | f | (200,90,150, result_test1, result_test2, med1, med2) | … | (200,90, 150, result_test1, result_test2, med1, med2) | … | (190,50,150,result_test1, result_test2, med1, med2) | Disease_Notcured |
| | 3 | 55 | m | (190,50,150,result_test1, result_test2, med1) | … | (200,90, 150, result_test1, result_test2, med1, med2) | NA | NA | Disease_cured |
| | 4 | 35 | m | (200,90,150, result_test1, result_test2, med1, med2) | … | (200,90, 150, result_test1, result_test2, med1, med2) | … | (190,50,150,result_test1, result_test2, med1, med2) | Disease_Notcured |

## IV. Sequence Mining using Weighted Association Rule

### a) Data Preparation

Data preparation process includes preparation of the data of interest to be used for mining and convert this data to the format suitable to perform apriori type algorithm. The database of the form shown in Table I have to be converted into the form as shown in Table IV.

#### i. Discretisation/ Normalisation

In the database firstly we perform Discretisation/ Normalisation for the non temporal attributes like age, gender etc. Discretisation is the process of converting the range of possible values associated with a continuous data item (e.g. a double precision number) into a number of sub-ranges each identified by a unique integer label; and converting all the values associated with instances of this data item to the corresponding integer labels. For example for attribute age the sub-range can be as shown in Table II

*Table 2 :* Discretisation Of Numeric Attribute

| age | categorical value |
|---|---|
| 20-30 | 1 |
| 31-40 | 2 |
| 41-50 | 3 |
| 51-60 | 4 |

*Normalisation* is the process of converting values associated with nominal data items so that they correspond to unique integer labels. TableIII shows normalization for attribute gender.

*Table 3 :* Normalization Of Attribute Gender

| Gender | integer label |
|---|---|
| male | 5 |
| female | 6 |

We use *DN (discretization/ normalisation) software Version 2* available at site http://cgi. Csc.liv.ac.uk/~frans/ KDD/ Software/ LUCS -KDD -DN/exmpleDnnotes. html to perform Discreti sation/ Normalisation process.

### *b) Similarity Search for Sequences using Euclidean Distance*

This is an important step in this framework. As once the database has been converted from database of sequences to the database of items, the apriori algorithm can be applied to find the association among the items, and ultimately the CAR rules can be generated for prediction.

To assign a unique numeric label to every unique sequences corresponding to each patient, sequence comparison method is required. There can be number of methods to compare the similarity of sequence.

Many time series representations and distance measure techniques have been proposed for more than one decade. Some of these approaches work well for short time series data, but they fail to produce satisfactory results for long sequences. There are two kinds of similarities: shape-based similarity and structure-based similarity. Shape based similarity is suitable for short sequences only. For the two sequences Si and Sj, shape-based determines the similarity based on local comparisons.

The well-known distance measure in data mining is Euclidean distance, in which sequences are aligned in the point-to-point fashion, i.e. the ith point in sequence Si is matched with the ith point in sequence Sj. Euclidean distance works well in many cases. Dynamic Time Warping (DTW) is another distance measure technique that overcomes the limitation by determining the best alignment to produce the optimal distance. Euclidean distance is a special case of DTW, where no warping is allowed, the dips and peaks in the sequences are miss-aligned and therefore not matched. In DTW, the dips and peaks of sequences are aligned and it provides more robust distance measure than Euclidean distance, compensation to that DTW a lot more computationally intensive as discussed in [13].

To determine the similarity for long sequence a more appropriate is to measure their similarity based on higher-level structures. Several methods for structure or model-based similarities have been proposed.

In this paper we use Euclidean distance measure for similarity search, For matching sequences we would like to address the following points.

- The relative times that the corresponding samples are collected are almost same in both sequences. This means that the lengths of sequences should be close to each other to be matched.
- The elements of both sequences are taken from the lifetime of the experiment in a rather uniformly manner.
- In numeric sequences from medical domains, since the elements are real numbers obtained from various pathological tests with a limited precision, elements from different sequences should be matched based on proximity.
- In non-numeric sequences, matching is done based on equality of their domain.

*Definition 6: Sequence Similarity:* Consider two sequences $S_1$ and $S_2$ having length x and y respectively, and $e_1, e_2 \ldots e_x$ are matching $q_1, q_2, \ldots . q_y$.

1. The sequences S1 and S2 matches each other if-
   i. Their length is same as

   ie $length(S_i) = length(S_j)$

   ii. Distance $(e_k, q_k) = 0$ ,for all values of k.

Also the distance between two elements ek and qk can be defined as follows.

- For numeric elements,

  distance $(e_k, q_k) = |e_k - q_k|$.

- Non-numeric sequences can be matched based on equality. In that case, the distance between any two elements is defined to be

$$\mathrm{distance}(e_k, q_k) = \begin{cases} 0 & , \text{if } \mathrm{dom}(e_k) = \mathrm{dom}(q_k) \\ \text{positive number}, & \text{if } \mathrm{dom}(e_k) \neq \mathrm{dom}(q_k) \end{cases}$$

2. and $S_i \neq S_j$ if either condition i or ii is false.

### *c) Representation of Temporal Sequences*

In order to perform the apriori like operation in the above dataset, we transform the original dataset consisting of sequences into the relational database consisting of numeric labels like 1, 2, 3…..etc, where each numeric label represents unique sequence. Sequences in one record are compared for their similarity and unique symbol is assigned to unique sequence.

In the database D consisting of m columns and n rows, we precede row wise from top to bottom and in each row we will precede from left to right. A unique numeric label num is assigned to the first sequence S(0,0) of first patient and maintains the processed sequence and num assigned to that sequence in arr as shown in Table V. Then we pick the next sequence S(I,j)and compare (using Euclidean distance) it with already processed sequences stored in arr. If the sequence matching is found then assign the same numeral to new sequence and no need to assign new numeric label. If the sequence is not present in the List arr then assign a num++ to the sequence and store the sequence and num to arr. Comparing the sequence in the list will always starts from first entry in arr but it will not be a time consuming process as there will be finite number of sequence in the original database D. This way entire database D is preprocessed to convert the database D' as shown in TableIV. The algorithm is discussed in figure 3.

*Table 4 :* Transformed Database D'

| Patient Record→ | Time Dimension→ | | | $t_1$ | … | $t_i$ | ..<br>. | $t_m$ | |
|---|---|---|---|---|---|---|---|---|---|
| | P_Id | age | gender | $s_1$ | ..<br>. | $s_i$ | … | $S_m$ | Class_ label |
| | 101 | 2 | 5 | 7 | ..<br>. | 8 | … | 8 | 10 |
| | 102 | 1 | 5 | 8 | ..<br>. | 8 | … | 9 | 11 |
| | 103 | 4 | 6 | 7 | ..<br>. | 8 | 0 | 0 | 10 |
| | 104 | 3 | 6 | 8 | ..<br>. | 8 | … | 9 | 11 |

*Table 5* : List Consisting of Sequences and Numeric Labels

| Sequence | Numeric labels |
|---|---|
| $S_{(0,0)}$ | 7 |
| $S_{(0,1)}$ | 8 |
| $S_{(0,5)}$ | 9 |
| - | - |
| - | - |
| $S_{(n,m)}$ | 30 |

```
Initialize num with available numeric value
Initialize k=0;
for each i=0 to n-1
{for  each j=0 to m-1
{ exists=false
for each l= 0 to k
{ if arr[l,0]=D[i,j]
exists=true
D[i,j]=arr[l,1] } }
If not exist then
{arr[k,0]=D[l,j]
arr[k,1]=num;
D[i,j]=num
k++ and num++} }
```

*Figure 2 :* Algorithm to convert D to D'

*d) Assigning Weight to the Sequences using Maximum Likelihood Estimation.*

The weighted concept is used to improve the performance in terms of accuracy and number of rules generating as mentioned in [18]. In this paper the weighed concept have been utilized to assign more weights to the sequence (pathological test and medication to the patient at particular time period) having much impact on treatment of patient. Attribute is assigned weight based on the domain. For example item in supermarket can be assigned weight based on the profit on per unit sale of an item. In medical domain, symptoms can be assigned weight based on their significance on prediction capability. Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. By using iterative technique, the maximum likelihood estimator is measured upon varying probability values of items in the training dataset.

*e) Frequent Sequence Mining*

The problem of sequence mining has now been converted to frequent itemset mining in the database D' where items are nothing but sequence represented by numeric labels. Hence the following section contains terms and basic concepts to define sequence weight, sequencesetweight, recordweight, weighted support and weighted confidence for weighted associative classifiers.

The transformed training dataset D' consists of n distinct set of records i.e. D'= {$r_1$, $r_2$, $r_3$.... $r_n$}. Where each record is collection of varying number of labels (representing temporal sequence) and value of class label. Each record has unique identifier called PID.

*Definition 7:* Sequence weight It is same as Item weight in WARM[19]. In this work we have extended the definition for the sequences. Each sequence $S_i$ is assigned weight wi, denoted by w($S_i$), where $0<w_i<=1$. Weight is used to illustrate the significance of the sequence. Attribute is assigned weight based on the domain. For example item in supermarket can be assigned weight based on the profit on per unit sale of an item. In medical domain, symptoms can be assigned weight based on their significance on prediction capability. Weight is calculated from training data using maximum likelihood estimation and denoted by wi. Table V shows the synthetic weight assigned to the sequences.

*Table 5 :* Synthetic Weight Assigned To The Sequences

| S.No | Numeric labels | *Sequence* | Weight |
|---|---|---|---|
| 1 | *7* | *(190,50,150,result_test1, result_test2, med1)* | 0.5 |
| 2 | *8* | *(200,90, 150, result_test1, result_test2, med1, med2)* | 0.6 |
| 3 | *9* | *(190,50,150,result_test1, result_test2, med1, med2)* | 0.8 |

*Definition 8 :* Sequence Set Weight: It is same as Itemset weight in WARM[19]. In this work we have extended the definition for the sequences set weight. Weight of sequence set X is denoted by W(X) and is calculated as the average of weights of enclosing attribute. And is given by

$$W(X) = \frac{\sum_{i=1}^{|X|} W(S_i)}{\text{Number of sequences in } X}$$

*Definition 9 :* Record Weight: The tuple weight or record weight can be defined as type of sequence set weight. It is average weight of sequences in the patient record. If the transactional table is having m number of sequence then Record Weight is denoted by W(rk) and given by

$$W(r_k) = \frac{\sum_{i=1}^{|r_k|} W(S_i)}{m}$$

*Definition 10 :* Weighted Support: In associative classification rule mining, the classification association rules R: X→Y is special case of association rule where Y is the class label. Weighted support (WSP) of rule X→Class_label, where X is non empty set of sequences, is fraction of weight of the record that contain above sequence set relative to the weight of all transactions. This can be given as

Here *m* is the total number of records.

*Example:* Let sequence Si= (190,50, 150,result_test1, result_test2, med1) and

Sj= (200,90, 150, result_test1, result_test2, med1,med2)

Consider a rule R: (( Si, Sj) → Class_label) then Weighted Support of R is calculated as:

$$WSP(X \rightarrow Class\_label\ ) = \frac{\sum_{i=1}^{|X|} W(r_i)}{\sum_{k=1}^{|n|} W(r_k)}$$

*Definition 11 :* A frequent sequence is a set of sequences whose support is greater or equal than a user-specified threshold called minimum weighted support (WMin_sup). Given a dataset and WMin_sup, the goal of sequence mining is to determine in the dataset all the frequent sequences set whose support are greater than or equal to WMin_sup.

*Definition 12 :* Weighted Confidence: Weighted Confidence (WC) of a rule X→Y where Y represents the Class label and can be defined as the ratio of Weighted Support of (X→Y) and the Weighted Support of (X).

$$WC(X \rightarrow Y) = \frac{WSP\ (X \rightarrow Y)}{WSP\ (X)}$$

*f) Classification Association Rule Generation*

After generating all frequent item sets CAR rules are filtered using frequent item set having one of the class labels. Frequent item sets that does not contain any of the class label has to be removed. To generate significant CAR rules the weighted confidence threshold is used. Using WAC algorithm the set of CAR rules are generated as shown in figure 4. Finally the numeric labels are replaced by corresponding sequence using arr database and frequent sequences are generated as shown in figure 5.

*(5, 8, 10, 12)→Class_label1*
*(9, 13, 15, 17)→Class_label2*
*(12, 19, 22, 25)→Class_label2*

*Figure 3 :* CAR Rules Consisting Of Numeric Labels

$( S_{(0,2}, S_{(1,8)}, S_{(3,9)}, S_{(5,6)} ) \rightarrow Class\_label1$
$( S_{(91)}, S_{(8,2)}, S_{(91,5)}, S_{(61,7)} ) \rightarrow Class\_label2$
$( S_{(81,2)}, S_{( 121,9)}, S_{(32,2)} , S_{(202,5 )} ) \rightarrow Class\_label2$

*Figure 4 :* CAR rules consisting of sequences

```
ALGORITHM 1:WAC($D_{tr}$, $WMS$, $WMin\_conf$, $W$, $D_{ts}$)
Input: 1 $D_{tr}$: Training data, 2 $D_{ts}$: Test Data, 3 $W_s$: Weighted support, $WC$: Weighted confidence,
Output: Weighted Class Association Rule Base(WCARB)
WCARB← WeightedAssociationRuleMiner($D_{tr}$, $W_s$, $W$, $WMin\_conf$)
WeightedAssociationRuleMiner ($D_{tr}$, $WMS$, $W$, $WMin\_conf$)
Apply Apriori type algo. using weighted support to find frequent attribute sets $a$ where $a=a_1,a_2...a \in D_{tr}$
for (all frequent itemset $a_i \in a$)
        if $a_i$ does not contain $c_i \in C$ (Set of Class labels)
remove $a_i$ from $a$
        else
                generate Rule $R_i=(a_i- c_i) \rightarrow c_i$
if(Weightedconfidence($R_i$)> $WMin\_conf$)
   Store $R_i$ to $WCARB$
BuildPredictiveModel($WCADB$, $D_{ts}$)
Sort Rules $R_i$ of $WCARB$ w.r.t. their $WMin\_conf$ and store CAR rule in Rule_Base
For each record $r_i \in D_{ts}$ predict class label using Rule-Base
Find Accuracy of the system
If (accuracy> minimum threshold )
The Model is  suitable for prediction
```

*Figure 5 :* The WAC Algorithm

## V. Experiments and Results

We present the Temporal WAC results on real data collections of blood cancer disease.

### a) *Representation and Modeling*

The data has been collected for 30 patient. The database consists of maximum 5 time stamp as shown in figure 7, two Class label with the values –cure and not cure.

Structure of sequence is-

{Cancerous cell% , Therapy, Medicine(s) }

Example of few sequences available in the dataset are-

- { 30% , Chemotherapy, Zofran, Busulphan, Kadian}
- {40%, Raditiontherapy, Aclarbicin, Azacitidine}
- {30%, Immunotherapy, Adriamycin IV, Elspar inj}
- {20%, Targeted terapy, Nilotinib, GastroMARK}
- {92%, StemCelltransplantation, Aclarbicin, Photofrin}

In the above sequence, first element is the percentage of cancerous cell, second element is therapy given to the patient and rest elements are medicines.

The Experiments have been performed step by step following the framework shown in figure1.

| P_Id | Age | Gender | Time_slot1 | Time_slot2 | Time_slot3 | Time_slot4 | Time_slot5 | Class_Label |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |

*Figure 6 :* Schema of cancer dataset

### b) *Similarity Measure (Pre-Processing)*

Euclidian distance measure is used to convert the database consisting of above sequence to database consisting of numeric labels.

Total 26 unique sequences have been identified and 27, 28 are assigned as the numeric labels for "cure" and "not cure" class labels respectively.

### c) *Mining Operation*

The WAC algorithm shown in figure 4 is used to mine the database and CAR rule are generated for the different support value. With the CAR rules the accuracy is calculated using same training data and the result is shown in Table 5.

*Table 6 :* Car Rules Consisting Of Numeric Labels

| S. No | Support Value | CAR rules | Confidence | Accuracy |
|---|---|---|---|---|
| 1 | 0.15 | 16,20,21 →28 | 100% | 66% |
| 2 | 0.20 | 19,22→27<br>8,23→27<br>7, 16→28 | 100%<br>100%<br>80%<br>100% | 80% |
| | | 16,20→28 | | |
| 3 | 0.25 | 16,20→28 | 100% | 66% |
| 4 | 0.30 | 22→27<br>17→27<br>20→28 | 100% | 90% |

With the result shown in table 5 we conclude that accuracy is better in case of having CAR rules for all the class labels. The reason of less accuracy in case of single CAR rule may be the default class label we are assigning during accuracy calculations. The Efficient CAR rules can be generated using enough training record. The purpose of this experiment is to show that Framework shown in figure1 is possible to implement and can generate useful result in medical domain can be used for the purpose stated in introduction section. The authors are confident enough that improved result will be obtained if the experiment were to be performed on real data with little or no modifications.

## VI. Conclusion and Future Work

This work presents a new foundational approach to mine frequent sequence using weighted associative classifiers whose core idea is to assign weights to the attributes depending upon their importance in predicting the class labels. The proposed model can be used as an alternative, computerized decision aid to assist physicians to find the sequence of treatment that can be given to the patient. The author feels confident in declaring that the framework is feasible one in the medical domain.

## References Références Referencias

1. Riccardo Bellazzi,Fulvia Ferrazzi, and Lucia Sacchi, Predictive data mining in clinical medicine: a focus on selected methods and applications, @2011 John Willey & Sons , Inc . Volume 00, Januar y / Februar y 2011.
2. Cláudia M. Antunes , and Arlindo L. Oliveira, Temporal Data Mining: an overview, Lecture Notes in Computer Science pp 1-15.
3. Stefano concaro, temporal emporal data mining for the analysis of healthcare data,.ph.d. Thesis (2006-2009).
4. Tu Bao Ho, Trong Dung Nguyen,Saori Kawasaki, Si Quang Le, Dung Duc Nguyen, Hideto Yokoi, Katsuhiko Takabayashi, Mining Hepatitis Data with Temporal Abstraction.
5. Luca Chittaro, Carlo Combi, Giampaolo Trapasso, Data Mining on Temporal Data:a Visual Approach and its Clinical Application to Hemodialysis, Journal of Visual Languages and Computing, vol. 14, no. 6, December 2003, pp. 591-620.
6. Weiqiang Lin, Mehmet A. Orgun, and Graham J. Williams, Mining Temporal Patterns from Health Care Data.
7. Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R. (2005) Temporal data mining for the quality assessment of hemodialysis services. Artificial Intelligence in Medicine 34:25–39.
8. Sacchi, L., Larizza, C., Combi, C., Bellazzi, R. (2007) Data mining with temporal abstractions: Learning rules from time series. Data Mining Knowledge Discovery 15, 217–247.
9. Concaro, S., Sacchi, L., Cerra, C., Fratino, P., Bellazzi, R. (2008) Temporal Data Mining for the Analysis of Administrative Healthcare Data. In Proceedings of IDAMAP 2008 Workshop, Washington, 75–80, http://labmedinfo.org/download/lmi503.pdf.
10. Stefano CONCARO , Lucia SACCHI , Carlo CERRA , Riccardo BELLAZZI, Mining Administrative and Clinical Diabetes Data with Temporal Association Rules, Medical Informatics in a United and Healthy Europe, IOS Press, 2009 European Federation for Medical Informatics pp574-578.
11. Stefano Concaro, MS, Lucia Sacchi, Carlo Cerra, Mario Stefanelli, Temporal Data Mining for the Assessment of the Costs Related to Diabetes Mellitus Pharmacological Treatment, AMIA 2009 Symposium Proceedings Page – 119-123.
12. Stefano Concaro, Lucia Sacchi, Carlo Cerra, Pietro Fratino, and Riccardo Bellazzi, Mining Healthcare Data with Temporal Association Rules: Improvements and Assessment for a Practical Use, AIME 2009, LNAI 5651, pp. 16–25, Springer-Verlag Berlin Heidelberg 2009.
13. Matching and Indexing Sequences of Different Lengths, Tolga Bozkaya Nasser Yazdani Meral ¨Ozsoyoˇglu.
14. SRIVATSAN LAXMAN and P S SASTRY A survey of temporal data mining , Vol. 31, April 2006, pp. 173–198.
15. B. Liu, W. Hsu, and Y. Ma. “Integrating classification and association rule mining”, In KDD’98, New York, NY, Aug.1998.
16. W. Li, J. Han, and J. Pei. “CMAR: Accurate and efficient classification based on multiple class-association rules” In ICDM'01, pp. 369-376, San Jose, CA, Nov.2001.
17. Yin, X. & Han, J. “CPAR: Classification based on predictive association rule”, In Proceedings of the SIAM International Conference on Data Mining. San Francisco, CA: SIAM Press, pp. 369–376, 2003.
18. Sunita Soni , O.P.Vyas, Performance Evaluation of Weighted Associative Classifier in Health Care Data Mining and Building Fuzzy Weighted Associative Classifier, D. Nagamalai, E. Renault, and M. Dhanushkodi (Eds.): PDCTA 2011, CCIS 203, pp. 224–237, 2011.© Springer-Verlag Berlin Heidelberg 2011.
19. Feng Tao, Fionn Murtagh and Mohsen Farid. “Weighted Association Rule Mining using Weighted Support and Significance Framework”, In proceedings of the ninth ACM SIGKDD International conference on Knowledge Discovery and Data Mining 2003, pp:661-666.

This page is intentionally left blank

Global Journals Inc.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Identification of Critical Risk Phase in Commercial-off-the-Shelf Software (CBSD) using FMEA Approach

By Palak Arora & Harshpreet Singh

*Lovely Professional University, India*

*Abstract-* COTS based development is becoming a popular software development approach for building large organizational software using existing developed components. COTS based approach provides pre-developed components either as in house or commercial off the shelf components, which reduces effort and cost for developing the software. There are potential challenges, risks and complexities in using COTS components. This paper provides an analysis of risks and challenges faced during developing software using CBSD approach. The risks under various phases are identified, categorized and prioritized the risks in various phases of CBSD and provide the mitigation strategy to manage the risks.

*Keywords: CBSD, risks in CBSD, risk mitigation.*

*GJCST-C Classification: K.6.3*

# Identification of Critical Risk Phase in Commercial-off-the-Shelf Software (CBSD) using FMEA Approach

Palak Arora [α] & Harshpreet Singh [σ]

*Abstract-* COTS based development is becoming a popular software development approach for building large organizational software using existing developed components. COTS based approach provides pre-developed components either as in house or commercial off the shelf components, which reduces effort and cost for developing the software. There are potential challenges, risks and complexities in using COTS components. This paper provides an analysis of risks and challenges faced during developing software using CBSD approach. The risks under various phases are identified, categorized and prioritized the risks in various phases of CBSD and provide the mitigation strategy to manage the risks.

*General Term: commercial-off-the-shelf software development (CBSD).*

*Keywords: CBSD, risks in CBSD, risk mitigation.*

## I. Introduction

COTS-based software development aims in building the software using the existing developed components. The components can be developed in house for usage among vast projects of similar requirements. The components can also be purchased from the market as the components are also developed as small software's which intend to provide the basic functionality required for large projects.

Various components are also available in the repositories with their functionalities and Quality attributes. A target application/ software are developed by selecting the appropriate components from the component repository & then integrating the components into a target system as in Figure 1 below.

At present time, more than 60% of software are developed using component approach due to its enormous features such as:

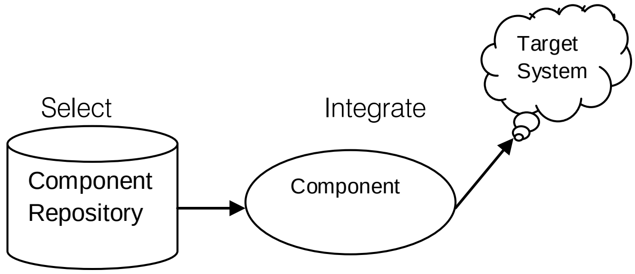


*Figure 1 :* Component-based Software Development

- Rapidly development.
- Accessed Immediately.
- Reduced Complexity.
- Increases efficiency of products.
- Reduced implementation, operating and maintenance cost.
- Reduced amount of time to deliver products in the market, budget and schedule saving, more than half of the software developers used component based approach. This approach has reduced the software crisis at great extent [6].

The main rationale of CBSD approach is to develop big system by integrating the pre-built components which decrease the progress time & costs. There are five main phases: Identification, Evaluation, Selection, Integration and Development of component to develop software using CBSD approach as mentioned in Figure 2 below.

## II. Review of Literature

To provide a reliable and effective software product in the market, software industry influenced by COTS development approach. In software applications CBSD is the only need to be written once and re-used multiple times than being re-written every time when a new application is developed. CBSD approach overlaps the traditional software engineering approach where existing technologies were failed to deliver project on-time and on-budget. The main reasons of these failures are: Testing -

*Author α: Student, School of CSE, Lovely Professional University Phagwara, Punjab. e-mail: palakarora718@gmail.com*

*Author σ: Assistant Professor, School of CSE Lovely Professional University Phagwara, Punjab. e-mail: harshpreet.17478@lpu.co.in*

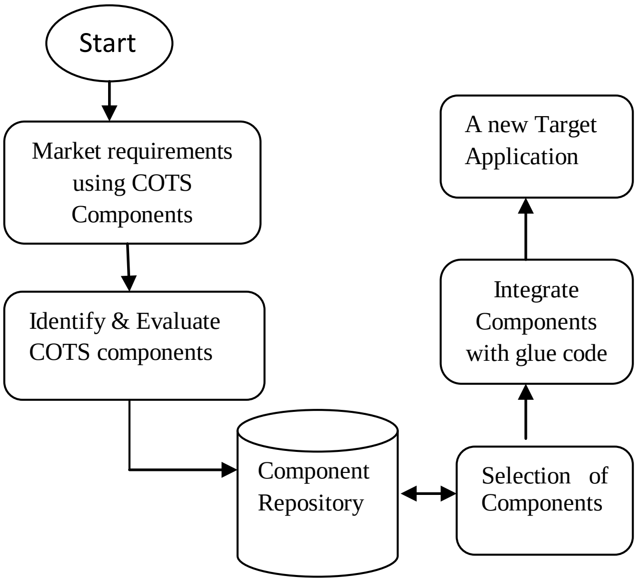

Start
Market requirements using COTS Components
Identify & Evaluate COTS components
Component Repository
Selection of Components
Integrate Components with glue code
A new Target Application

*Figure 2 :* COTS Development Life cycle

-efforts are not properly estimated; Team's skill is under/over estimated. However, the use of CBSD approach provides a lot of benefits, but still there are several challenges, risks, uncertainties related to this approach [6]. As the name suggested, CBSD approach means use of existing components, we are depending upon someone else (lack of trust). The main reasons of these problems are due to these factors:

- Wrong selection of components,
- Black box nature (non-availability of code) of COTS Components,
- Lack of knowledge, guidance etc.
- Unknown quality of COTS Products.

Many times, some risks are not identified in one phase and it overlaps to the second phase so in this way, it influences the whole software and fails to the organization's business. So, there is a need of proper Risk Management for using this CBSD approach from the starting phase. Failure Modes and Effects Analysis (FMEA) is a systematic method for evaluating a process to identify where risk is and how it might fail and to assess the relative impact of different failures [7]. With the help of FMEA approach, this paper provides risk management strategy for Commercial-off- The- Shelf Software development.

## III. Problem Definition & Solution

In developing software using CBSD approach there is an uncertainty that there can be variations between the planned development approach and the actual software developed. A risk could cause an organization to fail to meet its approach and objectives. The main steps of this paper are as in Figure 3 below:

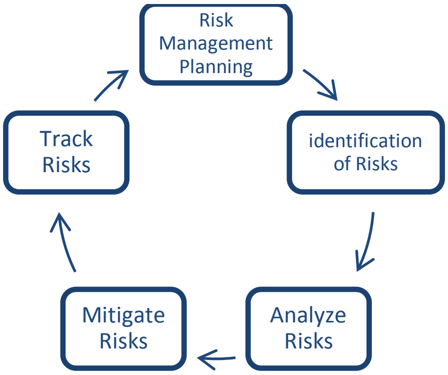


*Figure 3 :* Step-wise Problem definition

### a) *Challenges Faced during COTS-based Software Development life cycle*

The use of commercial-off-The Shelf software Development has become an important need for developing software as they offer reduce development time and effort. Similarly there are many challenges faced such as the quality attribute of selected components may cause deviation in the quality of final product, also the cost and effort involved in integrating component during the design process may cause the product design to deviate from the actual requirement There are many challenges that start during COTS development (Identification, Selection, Evaluation, Integration, and Development) summarised as below [1]: -

1. Companies have very limited access to product's internal design and the description of commercial package is in improper format.
2. When evaluating the COTS components, customers have very few chances to verify in advance whether the desired requirements will be met in the future.
3. Selection of COTS becomes major challenge faced by requirement engineers to match the requirements with available COTS.
4. Selection of components becomes major challenge faced by requirement engineers to match the requirements with available COTS.
5. The level of quality is unknown. The COTS products will have defects, no one know where and how many will be.
6. Documentation related to component may be of inadequate quality to be used.
7. Selection of COTS components is often based on subjective judgement, so there are no additional specifications provided by vendors for COTS component's internal architecture and description.

### b) *Risk Management Planning*

Risk management planning is a continuous process for identifying and measuring the risks continuously identifying and measuring the risks; developing mitigation options; selecting, planning, and implementing appropriate risk mitigations. It also involves tracking the implementation to ensure successful risk reduction.

i. *Identification of risks during CBSD Lifecycle*

Using the COTS development approach the components are purchased from the third party vendor due to which the development of the software depends upon the customer support services provided by the vendors. So, there are several chances of arising risks on each phase of CBSD as in figure 4. The risks in CBSD life cycle are due to the factors such as the black box nature of COTS components, lack of interoperability standards, the disparity between the user & suppliers, incomplete format of requirement documentation etc. The classification of risks based on various phases is briefly defined as in [6].

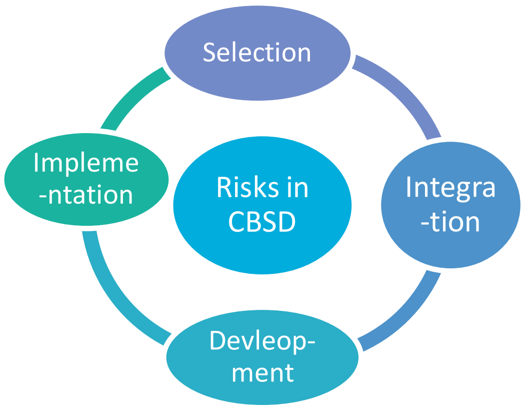


*Figure 4 :* Risks in CBSD life cycle

1. *Risks in COTS Selection Phase*

Risk during this phase is associated with the problems of evaluating and selecting off-the-shelf software for use in the system. The risks in this phase are due to some parameters as unavailability of source code, inflexibility of COTS components, lack of requirement document, architecture mismatches etc.

2. *Risks in COTS Integration Phase*

These risks are associated with problems of integrating systems from the existing COTS components. These risks can occur while composing of COTS components due to the lack of interoperability standards, occurrence of incompatible format among different COTS components, incomplete format of requirements etc.

3. *Risks during COTS Development*

The risks in this phase are arises when we develop the architecture from the selected COTS components. The risk arises due to the problem of using an inappropriate development process.

4. *Risks during COTS Implementation Phase*

The risks in this phase are during when we implement the final systems after selecting the appropriate components. These risks are due to the unclear design assumptions, performance factors, and security factors.

ii. *Classification of Risks during Phase-wise of CBSD*

There are three types of areas where the identified risk arises mostly:

- Functional/ Operational Requirements - The risks are which arises with the functionality and performance of the system as perceived by its operators.
- Procedural approach - The risks that are related with the technical characteristics of COTS products.
- Production strategy - Those risks which are related with the vendor of the COTS product.

1. *Risks Involving in Functional/ Operational Requirements*

*Table 1 :* Risks Involving in Functional/ Operational Requirements

| For this Potential types of risks | Risks |
|---|---|
| Availability Risks | In the case of COTS components, it is difficult to predict that the available COTS component will meet the functional requirements, so the estimated development cost and schedule are highly uncertain |
| Functionality & Performance | In COTS components, the actual functionality and performance of a COTS product are not as publicized so the system may not meet its requirements. |
| Requirements Gap | COTS component does not match the current operational requirements or procedures. |
| Security and Safety Issues | It may not be possible to certify that the product meets requirements because the COTS product must be tested as a black box without its implementation |

2. *Risk involving in Procedural Approach*

*Table 2 :* Risk involving in Procedural Approach

| For the possible kinds of Risks are: | Risks |
|---|---|
| Conformance to Commercial Standards | COTS components do not conform to commercial standards so interoperability with other selected COTS products may be difficult & costly. |
| Integration | Contractor does not have the technical |

| | |
|---|---|
| Contractor's Capability | Knowledge & experience to deliver a COTS-based system so the system may not meet requirements.. |
| Quality Requirements | COTS software components do not meet quality requirements (e.g., reliability, performance, usability). |
| Adaptability Risks | COTS products does not fully support initial and evolving requirements and do not have built-in flexibility. |
| Portability Risks | It is not necessary that COTS package will always supportable across a variety of hardware and operating system platforms, then hardware platform choices over a program lifecycle may be limited. |
| Evolution Risks | Sometimes, COTS components, hardware upgrades or replacements are not compatible with current COTS software products so COTS software components may have to be replaced at the same time. |
| Source code | If there is no access to source code, then it may be difficult to trace integration and testing problems to COTS products |
| Upgrades | Sometime during upgrading COTS software, it increases the size of the programs & the size of the hardware memory in the system may be insufficient. |

3. *Risks involving in Production Strategy*

*Table 3 :* Risks involving in Production Strategy

| For this potential kinds of Risks are: | Risks |
|---|---|
| Acquisition Alternatives Risks | During evaluation time, alternative methods of acquiring COTS products are not evaluated |
| Vendor Reliability Risks | Sometimes, the vendor of COTS product is financially weak or unstable & poor support. |
| Cost and Schedule Completeness: | The cost and schedule estimates are not considered during acquiring the COTS-based system. |
| Business Skills | The relationship between the contractor and vendor contractor are weak. |

iii. *Risk Mitigation*

The main focus is to track, control and reduce the identified risk. A survey was conducted in various CMM level 2 companies which summarized the possibility of risk and corresponding impact of risks. Two approaches are used to calculate the risk score of identified risks in order to plan mitigation approach for the high impact risks.

a. Failure Mode and Effect Analysis (FMEA)
b. Goal-Driven software Risk Management (GSRM)

a. *Failure Mode and Effect Analysis*

A failure mode and effects analysis (FMEA) is a method for examine of potential failure modes within a system for classification by the probability and likelihood of the failures [5]. This procedure helps a team to identify potential failure modes based on past experience with similar products, enabling the team to design those failures out of the system with the minimum effort and resource expenditure. Effects analysis refers to studying the consequences of those failures. To calculate the risk score of identified risks, we are using this approach & filled the questionnaire from the 12 team member based on their past experience of using COTS components.

The probability of each risk item is measuring on likert scale ranging from low (1), moderate (3), and critical (5) as below:

| Likert Scale | Probability measurement |
|---|---|
| Low | 1 |
| Moderate | 3 |
| critical | 5 |

The impact of corresponding risk item is ranging from very low (0) to critical (5) as below:

| Likert Scale | Impact values |
|---|---|
| Very low | 0 |
| Low | 1 |
| Moderate | 2 |
| High | 3 |
| Very high | 4 |
| Critical | 5 |

Here are some assumptions of choosing these values:

- It is assuming that the impact of each risk could be different at each phase; it could be or not be same at each phase.

- Suppose there is a probability of arising risk is Low (1), but its impact may be moderate (2) or may be critical (5).

The working formula is:

Risk Score= $\sum_{i,j=1}^{n} Pi * Ij$

Where Pi= Probability of Risk,

Ij = Impact of risk, n= number of respondents

*Results of questionnaire:* The results that have been conducted from the respondents are shown as below: -

1. *Risk score of Selection Phase*

*Table 4 :* Risk score of Selection Phase

| COTS Driver/Factor | Risk Id | Risk in Selection Phase | Risk Score |
|---|---|---|---|
| Behaviour Factors | RS1 | Unavailability of source code | 124 |
| | RS2 | Organizations have very limited access to product's internal design. | 108 |
| | RS 3 | The Quality level of a component is unknown. | 118 |
| | RS 4 | During evaluation, developers have limited chance to verify COTS behaviour. | 126 |
| Functionality Factors | RS 5 | Requirement of the user and component architecture does not match. | 174 |
| | RS6 | Architecture of the component is not analyzed according to the functionality. | 113 |
| | RS 7 | Difficult for requirement engineers to select among different techniques of selection. | 86 |
| | RS 8 | Lack of market survey. | 207 |
| Cost Factor | RS 9 | Required COTS is found costly as compared to in-house Development cost. | 69 |

Analysis of Risk Score

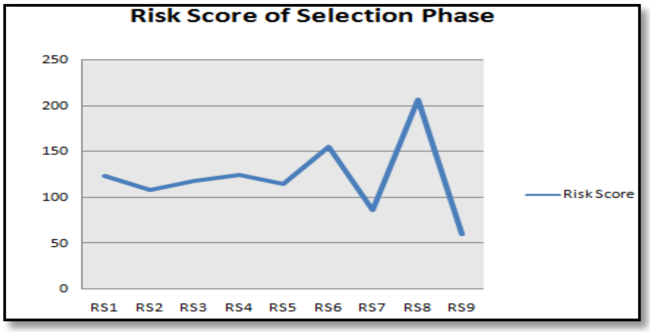


*Figure 4 :* Analysis of Selection Phase

From the above risk score, we analyzed RS5; RS 8 are critical risks because they have high impact of risks.

*2. Risk Score of Integration Phase*

*Table 5 :* Risk Score of Integration Phase

| Risk Driver/ Factors | Risk Id | Risks in Integration Phase | Risk Score |
|---|---|---|---|
| Cost Factors | RINT1 | Underestimate the development time and cost | 122 |
| | RINT2 | The cost is too much to configure the components | 83 |
| | RINT3 | Immature COTS components. | 91 |
| | RINT4 | Lack of requirement configurations. | 211 |
| | RINT5 | Lack of cost control. | 112 |
| Size Factors | RINT5 | Difficult to predict the size of components. | 132 |
| Personnel shortfall factors | RINT6 | Lack of knowledge. | 73 |
| | RINT7 | Lack of interoperability standard. | 146 |
| | RINT8 | Lack of integrator personnel. | 150 |
| Security factors | RINT9 | Vulnerability risks. | 140 |
| Functionality Factors | RINT10 | Unavailability of source code. | 137 |
| | RINT11 | Components are not platform independent. | 86 |

Analysis of Risk Score

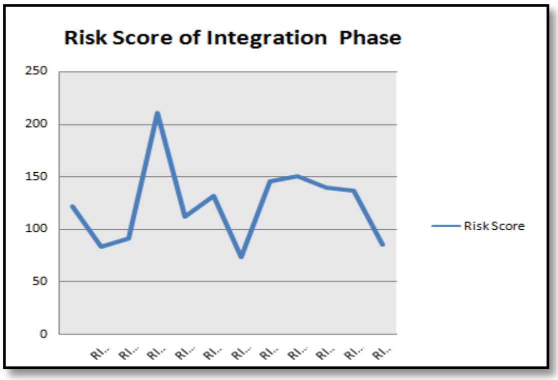


*Figure 5 :* Analysis of Integration Phase

From the above risk score of Integration phase, we analyzed that RINT 4, RINT 9 are critical risk; because they have high impact of risks.

ii. *Risk Score of Development Phase*

*Table 6 :* Risks Score in Development Phase

| Risk Drivers/ Factors | Risk Id | Risks in Development Phase | Risk Score |
|---|---|---|---|
| Inappropriate Development | RD 1 | Risk analysis phase is not present in CBSD. | 151 |

| | | | |
|---|---|---|---|
| Process | RD 2 | Risks are associated due to using an inappropriate development process. | 77 |
| Functionality Factors | RD 3 | A new version of COTS software may lack new updated code | 144 |
| | RD 4 | Resources are insufficient. | 106 |
| | RD 5 | Components are not properly supported by the vendor. | 148 |
| Behaviour Factors | RD 6 | The estimation of resources {time, cost} is exceeded during development for many projects. | 95 |

Analysis of Risk Score

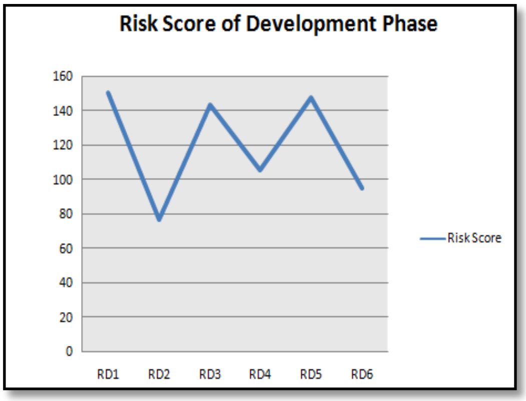


*Figure 6 :* Analysis of Development Phase

From the above risk score of Development phase, we analyzed that RD 1, RD 5 are critical risk; because they have high impact of risks.

iii. *Risk Score of Implementation Phase*

*Table 7 :* Risk Score in Implementation Phase

| Risk Drivers/ Factors | Risk Id | Risks in Implementation *Phase* | Risk Score |
|---|---|---|---|
| Functionality Factors | RI 1 | Unclear design assumptions. | 139 |
| Usability Factors | RI 2 | Users cannot retrieve relevant & needed information. | 97 |
| Security Factors | RI 3 | System can be used in unintended way. | 132 |
| | RI 4 | Increase in vulnerability attack by integrating components with one another. | 160 |
| Performance Factors | RI 5 | Effect on system performance. | 114 |

Analysis of Risk Score

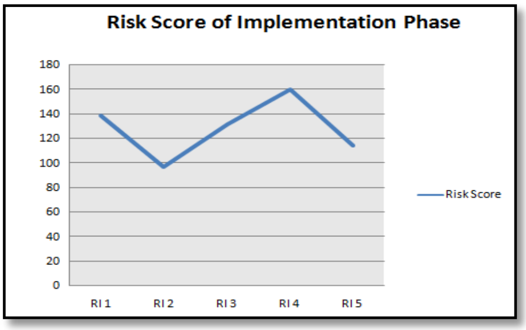


*Figure 7 :* Analysis of Implementation Phase

From the above risk score of Implementation Phase we analyzed that RI 1, RI 4 are critical risks because they have high impact of risks.

4. *Goal-Driven Software Risk Management (GSRM)*

During study it is analyzed that if the risk in one phase is unseen or undetected, it goes to the second phase and so in this way it impacts to the whole system. If the risk in one phase is not detected, it overlaps to the second phase and increases its multiplicative impact factor [5].

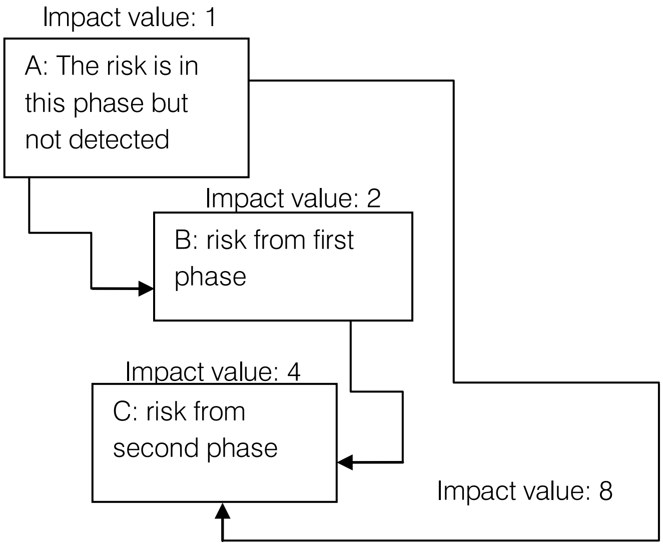


*Figure 8 :* Impact of Risks during phase-wise

In GSRM approach the main focus is to integrate the whole risk activities, so that we can identify those phases which have high impact of risks and then we can mitigate those risks. So we will calculate the total impact of risks as table 10.

The working formula to calculate total risk is as:

Total Risk Score= $\sum RS_k + \sum RINT_k + \sum RD_k + \sum RI_k$

Where $RS_k$ = Risk in Selection Phase,

$RINT_k$= Risk in Integration Phase,

$RD_k$ = Risk in Development Phase, $RI_{k=}$ Risk in Implementation Phase

i. *Total Risk Score of all CBSD (Commercial- Off-The- Shelf Development)*

*Table 8 :* Analysis of Total Risk Score

| Total impact of risk | |
|---|---|
| CBSD phase | Total Risk |
| Risk in Selection phase | 1098 |
| Risk in Implementation Phase | 1481 |
| Risk in Development Phase | 721 |
| Risk in Implementation Phase | 642 |

Analysis of Total Risk Score

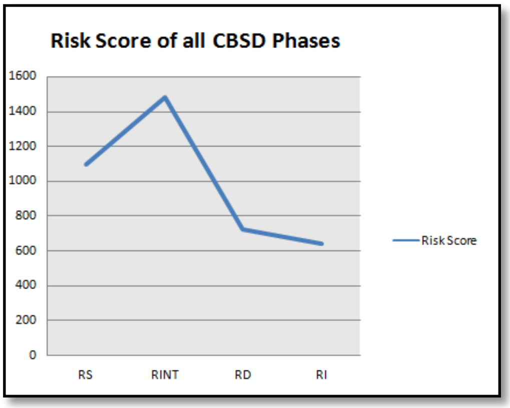


*Figure 9 :* Analysis of Total Risk Score

From the total risk score of all CBSD phases, we analyzed that Integration phase is more critical. So there is need to mitigate these risks.

*a) Risk Mitigation Strategy for Integration phase of CBSD Development approach*

From the results obtained during risk analysis, the following graph shows the risk score percentile in various COTS-based Development phases.

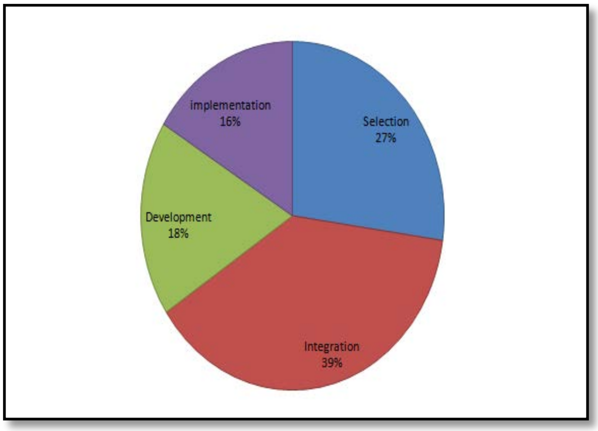


*Figure 10 :* Risk Score Percentile of all Phases

Now the mitigation strategy will be designed for most critical risk that is Integration Phase.

COTS Integration means when different COTS packages are combine into a system with “glue code”. For ex, Office Automation Software, email, messaging system, where the components are bundled as a procedural library [1]. But in this phase many risk arises as:

- Lack of interoperability standard.
- Lack of tools, methods to integrate components.
- Effort for integration may increase from what was estimated.
- When developers try to integrate incompatible COTS components etc.

This integration phase becomes a most challenging phase in Component-based Software Development. The main failures in software arise due to wrong integration of components. As in [4], the recent computer screen upgrade in the British Government caused nearly 80,000 desktop computers to crash The crash halted the United Kingdom’s pension and benefits agency that provides benefits to about 24 million people. The crash delayed the process of new claims and forced employees to fax and fill out some payment checks by hand. The problem occurred during an upgrade across the network of computers. So there is need to improve Integration techniques of COTS components.

Mitigation guidelines for Integration of COTS Components:

| | |
|---|---|
| 1. | A proper understanding of component’s capabilities is must how components are packaged and evaluated. |
| 2. | A developer should avoid general modifications to COTS components. |
| 3. | Modifications that add the complexity to the project of COTS components should be avoided. |

| |
|---|
| 4. When a developer add or replace a component, it should be integrated system testing. |
| 5. A proper documentation should be there before buying or developing components from third-party vendors. |
| 6. A developer should use the components that fulfil with well-known component standards. |
| 7. A developer, vendor or customer must have knowledge of integration tools. |
| 8. A developer should use reliable and trustworthy components so that it can minimise the risk of COTS system and provide quality to the system. |
| 9. The main risk in component system are due to the reason that components are not platform dependent with the system, a developer should provide components that supports adaption to the system |
| 10. While integrating the components, a developer should choose exact match of COTS components with system requirements instead of approximate match of COTS components. |
| 11. A developer should use open Standard technologies that are freely distributed among different data models or software infrastructure which provide basis for communication and enable consistency among different COTS components [6]. |
| 12. A proper estimation of time and cost should be estimated, before integrating COTS Components. |
| 13. All drivers should be considered before measuring component behaviour. For ex, ACIEP-used for COTS Integrator Experience with the product,<br>ACIPC - used for COTS Integrator Personnel Capability. |

## IV. Conlcusion

Commercial-off-The-Shelf Software Development has become a great need for large organizations as it saves development time and money. It is belief that COTS components fulfill everyone's needs and can be used as-is. In reality, the risk arises in each phase of CBSD as, COTS selection, Integration, Development and on maintenance phase. In this paper, the main focus is to provide risk identification strategy for COTS based software Development. The risk adds on each phase of CBSD was identified and risk score is calculated to examine the critical risk phase.

## References Références Referencias

1. Dr. Sohail Asghar, Mahrukh Umar, " Requirements Engineering Challenges in Development of Software Applications and selection of Customer-off-The-Shelf (COTS) components", in International Journal of Software Engineering(IJSE), 2010, (pp 32-50).
2. "Risk Management Guide for DOD Acquisition", in OUSD (AT&L) Systems and Software Engineering/Enterprise Development.
3. James Everett Tollerson, Hisham M. Haddad, "Conceptual Model for Integration of COTS Components" in Department of Computer science &IT (pp 1-7).
4. Amandeep Kaur & Shivani Goel, "Designing of RIMCOTS model for Risk identification and mitigation for COTS-based Software Development" in Research Journal of Computer Systems Engineering- an International Journal.
5. Saima Amber, Narmeen Shawoo & Saira Begum, "Determination of Risk During Requirement Engineering Process" in Journal of Emerging Trends in Computing and Information Sciences (pp 358-364).
6. Palak Arora, Amandeep kaur, "Improving COTS-based Software Development Process by Identification and Mitigation of Component Risks" in International Journal of Advanced Research in Computer Science and Software Engineering, 2013, (pp 219-225).
7. "Failure Effect Mode Analysis (FMEA) "in Institute for healthcare Improvements.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Effects of Mining Operations on Local Area Networks in Large Scale Gold Mining Environments in the Western Region of Ghana

By Emmanuel Effah & Christian Kwaku Amuzuvi

*University of Mines and Technology, Ghana*

*Abstract-* We investigate the impacts mining operations have on established Wired/Wireless Local Area Networks (WLANs) in mining environments in the Western Region of Ghana. Mining activities have certain immutable negative impacts on the topography of the land with consequent effects on LAN Networks. Notable are undulating landscape with pronounced physical obstructions, LAN infrastructural relocations and reconstruction, higher atmospheric dust concentration, severe ground vibrations due to blasting and the motion of heavy mine machineries. The mobile nature of mining operati-ons/practices often results in relocations of established network infrastructure such as fibre cables, repeater base stations, and mask towers (i.e. cell sites).The main reason for LAN infrastructural relocations is to ensure effective LAN/WLAN communication especially during mine expan-sions. However, this results into lengthy network downtimes. Employees' redundancies or idleness during network downtimes reduce mine productivity by about GHc2, 577, 860.64 (USD 1,288,930.32) annually. We recommend preventive maintenance schedule for all existing LAN infrastr-ucture; basic Information Communication Technology (ICT) Training into the regular training module; technically qualified Information Technology (IT) experts be part of management and finally; IT projects be planned and integrated into the annual business plan. Netronics Wireless Broadband (NWB) communication technology solutions were also recommended to management and IT policy makers in the mining companies for consideration due to its good performance in mining environments.

*Keywords: information communication technology, info-rmation technology, local area network, wired/wireless local area networks, intranets/extranets, infrastructural relocations.*

*GJCST-C Classification: C.2.5*

# Effects of Mining Operations on Local Area Networks in Large Scale Gold Mining Environments in the Western Region of Ghana

Emmanuel Effah [α] & Christian Kwaku Amuzuvi [σ]

*Abstract -* We investigate the impacts mining operations have on established Wired/Wireless Local Area Networks (WLANs) in mining environments in the Western Region of Ghana. Mining activities have certain immutable negative impacts on the topography of the land with consequent effects on LAN Networks. Notable are undulating landscape with pronounced physical obstructions, LAN infrastructural relocations and reconstruction, higher atmospheric dust concentration, severe ground vibrations due to blasting and the motion of heavy mine machineries. The mobile nature of mining operations/practices often results in relocations of established network infrastructure such as fibre cables, repeater base stations, and mask towers (i.e. cell sites).The main reason for LAN infrastructural relocations is to ensure effective LAN/WLAN communication especially during mine expansions. However, this results into lengthy network downtimes. Employees' redundancies or idleness during network downtimes reduce mine productivity by about GHc2, 577, 860.64 (USD 1,288,930.32) annually. We recommend preventive maintenance schedule for all existing LAN infrastructure; basic Information Communication Technology (ICT) Training into the regular training module; technically qualified Information Technology (IT) experts be part of management and finally; IT projects be planned and integrated into the annual business plan. Netronics Wireless Broadband (NWB) communication technology solutions were also recommended to management and IT policy makers in the mining companies for consideration due to its good performance in mining environments.

*Keywords: information communication technology, information technology, local area network, wired/wireless local area networks, intranets/extranets, infrastructural relocations.*

## I. Introduction

Enterprises depend on information which must be communicated accurately, securely, and quickly. This information is often created on a myriad of hardware and software platforms, thereby increasing the difficulty for its effective and efficient exchange [1]. These rapid developments in computer technology have resulted in a greater reliance on distributed computing, typified by "client/server" [2]. Again, the increasing reliance on networks driven by the growing use of sophisticated applications has created the desire for more faster and uninterruptible network or "backbone" - WLAN/LAN. Additionally, the influx of Intranets/Extranets and the Internet technologies coerce companies to building more resilience and guaranteed networks with much reduced downtimes so they can effectively survive competition. Earlier, Network failures were much routine and unplanned for which reason downtimes were measured in days. Today, networks unavailability for even a relatively short period of time cause substantial loss to the business.

Mining companies now keenly rely on LAN for sharing information, data, and technology resources, and completely show zero tolerance for network downtimes. Thus, the long held belief that 80% of traffic remains local to the network, while 20% traverses the backbone is no longer true. In fact, there has been nearly a total reversal in LAN traffic patterns now being called "20/80 rule" [1]. The prevalence of higher intensities of dust, severe noise and vibrations due to the use of various degrees of explosives, movement of heavy mine machineries and physical obstructions at most mining environments are detrimental to the effectiveness of LANs [3-6]. The nomadic nature of mining itself also create greater hindrance to LANs' efficiency (be it wired or wireless) [7]. Normal mining practice is that, as the ore at a place gets exhausted, mining activities must relocate and hence communication infrastructure must be moved. Consequently, laid fibre optic cables, transmitting/repeaters stations must be abandoned or relocated. Line-of-sight wireless signals is obliterated due to the abrupt topological changes in landscape and "kinking" of laid fibre optic cable create sustained network downtimes. The mobility of mining operations and the subsequent relocations of the installed LAN infrastructure and the peripheral devices, and even the cost of network reconstruction create a lot of inconveniencies. The extent to which these impede the Intranets' services demands attention, because the resulting accrued network downtime cost could be too huge. The reason being that, relocation of LAN infrastructure comes with its own demerits especially if unplanned [7-8]. Relocation technicalities are always impeded; thus, getting the required expertise, resources to do it and getting the desired material. Under-utilization of LAN due to frequent

*Author α σ : University of Mines and Technology, Department of Computer Science and Engineering/Department of Electrical and Electronic Engineering. Tarkwa, Ghana.*
*e-mails: ckamuzuvi2000@yahoo.com, ckamuzuvi@umat.edu.gh*

downtimes is more expensive to organizations than when efficiently utilized [9].

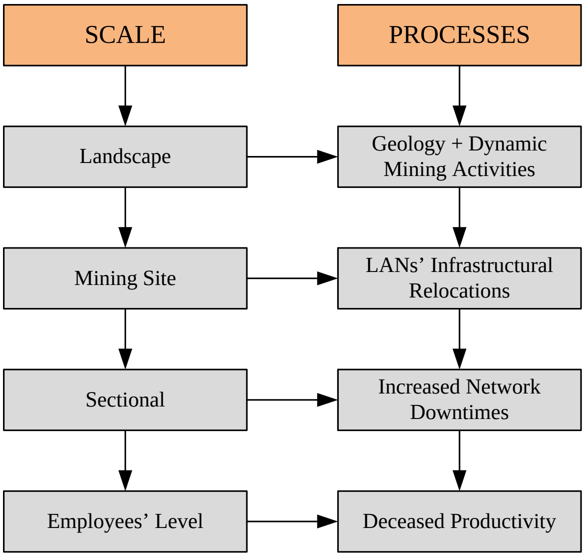

SCALE
PROCESSES
Landscape
Geology + Dynamic Mining Activities
Mining Site
LANs' Infrastructural Relocations
Sectional
Increased Network Downtimes
Employees' Level
Deceased Productivity

*Figure 1 :* Problem Structure

Mining operations are dynamic in nature. Lowlands are stockpiled to become highlands overnight and vice versa. Relocation of mine administrators' offices/workshops, mineral (gold) processing plants, fibre optic cables, repeater/transmission stations and human settlements or human communities are classic mining practices. Figure 1 presents the structure of the problem. Peculiar to this study is the way the mobile nature of mining itself and its consequent LAN infrastructural relocations affect network functionality and employee productivity. This research addresses this gap.

## II. Materials and Methods

This study deployed the descriptive research method involving observations and surveys [10]. Information about the existing condition was gathered using interviews, questionnaires and observations [11]. First hand data from the respondents was collected and analysed to form the basis for the conclusions and recommendations.

The research was limited to large scale gold mining companies within the western Region of Ghana and did not test any hypothesis or quantifiable data to generalize the results. Rather, this work sought thorough information and a deep understanding [12], of the stipulated research problem [13-14]. The qualitative research approach was therefore used.

## III. Results

The analysis and presentation of results were done in the order of the questionnaires viz: respondents' profile, Random LAN infrastructural relocations and LAN network effectiveness due to the mobility of mining operations, employees experience and response to network issues. The Statistical Package for the Social Sciences (SPSS) v16 and Microsoft Office Excel-2013 application software, were used in the analysis.

### *a) Demographic Profile of Respondents*

This part of the questionnaire looked at gender, departments, and work experience with their respective companies, rank and educational background.
From the survey, it was found out that 39.3% of the respondents were females and 69.7% males which are typical of gold mining companies. Figure 2 below illustrates the graphical distribution of employees in their various departments. Respondents solely relied on the installed LAN and it accessories to execute their daily duties as employees.

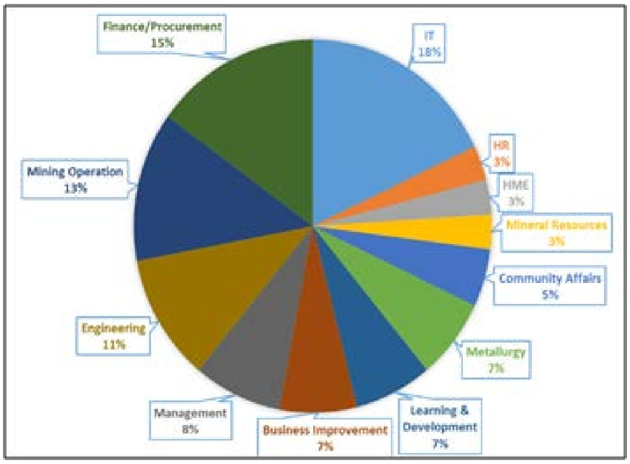


*Figure 2 :* Departmental distribution of respondents

Table 1 presents the Working Experience of Employees in the mine. Among the respondents interviewed, 63.9% were Senior Staffs or Managers (belonging to C3-C4 payment category), 34.4% people were supervisory staff (belonging to the C5-D1 payment category) and senior managers (belonging to the D2-D-upper payment category) representing 1.7%.

*Table 1 :* Years of work in the company

| Years in the Mines | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| Up to 5 | 46 | 75.4 | 75.4 | 75.4 |
| 11-15 | 2 | 3.3 | 3.3 | 78.7 |
| Over 21 | 2 | 3.3 | 3.3 | 82.0 |
| 6-10 | 11 | 18.0 | 18.0 | 100.0 |
| Total | 61 | 100.0 | 100.0 | |

Figure 3 presents respondents Educational levels. Regarding ICT training and qualifications 26% of the employees interviewed have formal ICT training with qualifications to that effect whereas 74% do not have.

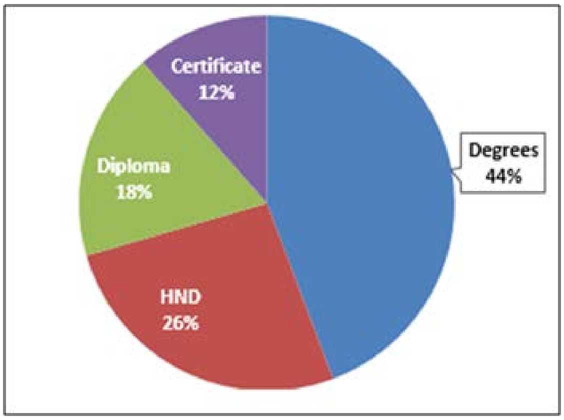


*Figure 3 :* Highest Educational levels of Respondents

### *b) Drivers for LAN Deployment*

Different organizations or companies deploy ICT for different purposes. In probing why mining companies were using LAN or ICT, respondents expressed their views depending on the kind of services their respective sections or departments receive from the LAN or Intranet. Figure 4 illustrates drivers for LAN services as deployed in the mine.

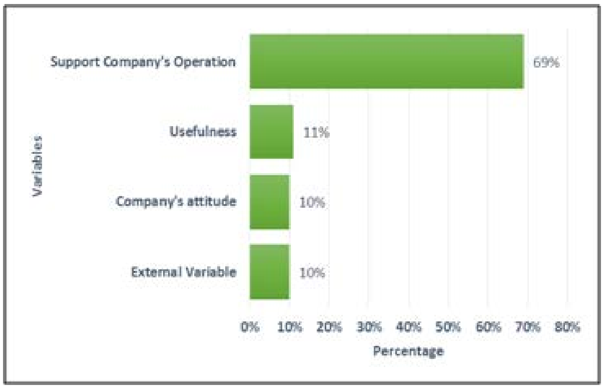


*Figure 4 :* Drivers for LAN Services Deployment in the Mine

### *c) LAN Infrastructural Relocations, Network Availability and Lost in Productivity due to Network Failures or Downtimes*

It is known that, "increased LAN/WLAN network infrastructural relocations resulting in LAN/WLAN network downtimes in mining operational environment decreases mine productivity". In order to affirm this fact, questionnaires administered ascertained the lost productive hours of employees as a result of network unavailability (downtimes) and employees' experience and response to network challenges.

In order to ensure certainty and establish good grounds for results, the extent of respondents' dependency on LAN link or the Intranet or the Internet in the daily basis was explored. Per this research, 93% of the employees confirmed sheer dependence on the LAN network link availability and completely became redundant if the link was down. Averagely, this value represents more than 900 employees for a mining company. 7% however, could execute their daily duties even when the network link was down.

Reasons and impacts of LANs' infrastructural relocation were to cater for expansion and improvement in network efficiency especially when well-planned and budgeted for. However, this study shows that the unplanned relocations surpass the planned. Figure 5 summarizes the root causes of LAN infrastructural relocations. As shown in Figure 6, almost half of the population (46%) believes LAN infrastructural relocations are means to expanding the network.

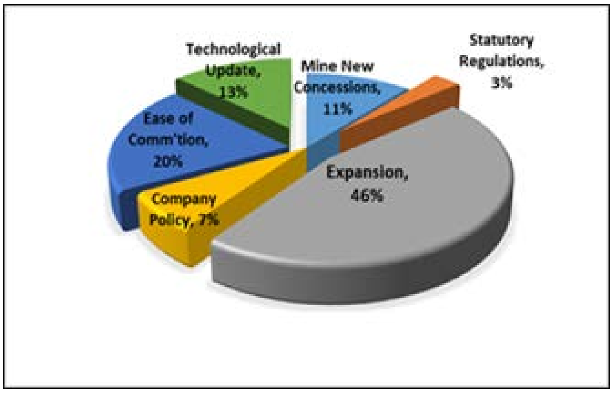


*Figure 5 :* Reasons for LAN Infrastructural Relocations

Figure 6 displays the impacts of LAN infrastructural relocations. Improving LAN's efficiency and minimizing interference due to noise, dust and stray frequencies from old sites are the intended impacts as subscribed by 52% of the respondents. However, the consequent reduction in LAN's efficiency due to prolonged link downtimes, increased network usability cost and maximized interference due to noise, dust and stray frequencies from new sites constitute the real impacts. 48% of the respondents alleged that the negative impacts surpass the positives.

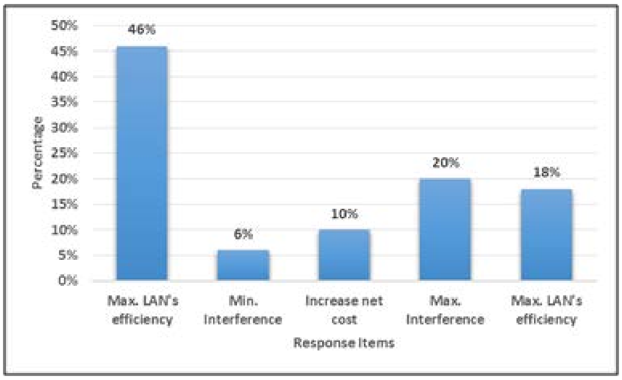


*Figure 6 :* Impacts of LAN Infrastructural Relocations

According to the respondents, the term "random" is frequently used to describe relocations because whenever newly explored concessions are to be mined, relocating LAN/WLAN infrastructure are considered minor tasks normally not well planned and factored into annual budgets. In fact, LAN/WLAN relocations are done to ease Internet and Intranet communication during expansion to mine new concessions. Actually, the major intended impacts of LAN infrastructural relocations on network function and availability are to improve LAN/WLAN's efficiency and minimize interference.

The realistic and inevitable repercussions of LAN infrastructural relocations on network function and availability according to 48% of the employees include:

- Increased network usability cost due to reworks during relocations and non-alignment with existing technology.
- Maximized interference (disturbance) due to noise, dust, space and other stray frequencies at new sites.
- Reduced LAN's efficiency and hence productivity due to prolong link downtimes.

From the analysis and the above deductions from employees, causes and reasons for LAN infrastructural relocations are logical. Nevertheless, their consequent impacts on network availability, effectiveness and hence mine productivity of network-using employees is negative.

### *d) Productive Hours Lost through Network Downtimes*

From Figure 7, employees experience rapid and sporadic network downtimes. 39% of the respondents see not less than one network downtime per day; 36% encounter not less than one network downtimes in two days; 25% experience network downtimes at least once a week. Establishing blameless baseline for logic analysis, we realized that, averagely, the network link goes down at least once in every two days.

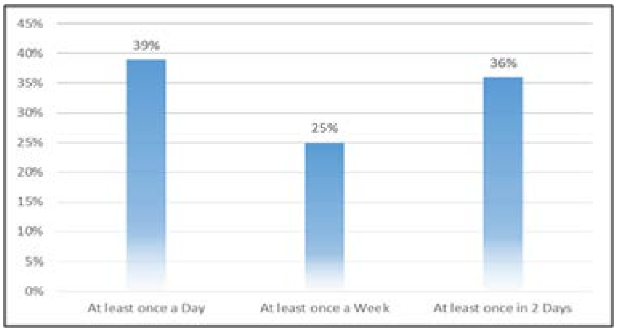


*Figure 7 :* Rate of Occurrences of LAN connectivity Problems

Figure 8 presents similar but at a broader perspective at the departmental level.

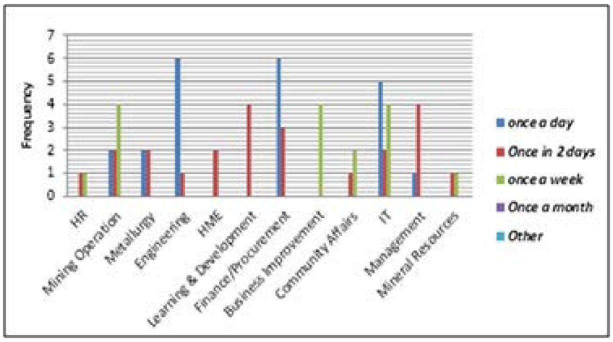


*Figure 8 :* Rate of LAN Network downtimes at the Departmental Level

The lengths of downtimes are illustrated in Table 2. Figure 8 shows how employees expend this time. As established from Figure 8, Table 2 extrapolates the length of downtimes averagely in two days per employee. From Table 2, 4.77 hours of productivity per an employee were lost every two days due to network downtimes. As broadly illustrated in Figure 9, more than 60% of the absolute LAN dependents waste over four productive hours every two days as a result of LAN network downtimes.

Figure10 shows that about 51% of the population do not channel their network challenges to IT desk, which can significantly delay network restoration. From Figure10, only 49% directly report their network grievances to the IT help desk.

*Table 2 :* Frequency Distribution of Network Downtimes

| Downtime ($x$) | Frequency ($f$) | $fx$ |
|---|---|---|
| 4 hours | 25 | 100 |
| 5 hours | 10 | 50 |
| 6 hours | 15 | 90 |
| 7 hours | 5 | 35 |
| 8 hours | 2 | 16 |
| Other | 4 | 0 |
| | $\sum f = 61$ | $\sum fx = 291$ |
| $Average\ Downtime = \frac{\sum fx}{\sum f} = \frac{291}{61} = 4.77 Hours$ | | |

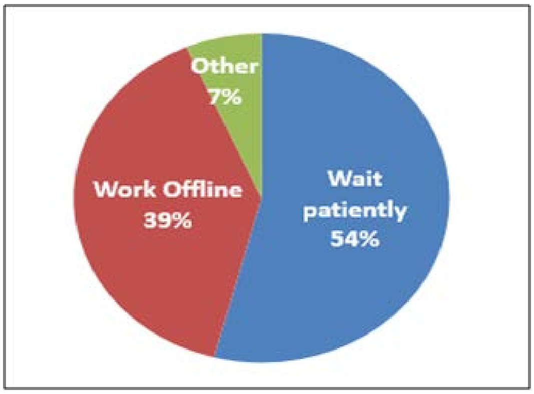


*Figure 9 :* What Employees do during Network Downtimes

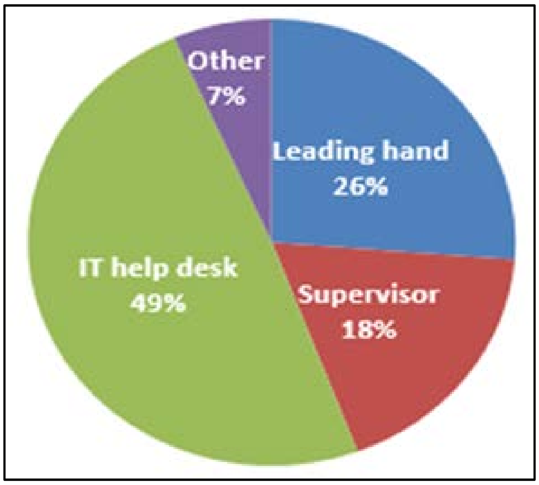


*Figure 10 :* Points of Contact during Network Failures or Issues

### e) *Downtime Deductions on Productivity or Business Operations*

For the twelve Intranet/Internet-using departments selected for this research, 97% of the respondents fully rely on the network to carry out all their daily operations while 3% can operate offline. This 97% represents over one thousand employees. Again, 74% of the respondents do not have any ICT training, be it formal and informal including management.

Alarmingly, 74% of the respondents experience network failures at least once in one or two days while 26% replied at least once a week. The most vital departments forming the core of production: engineering, finance and procurement, recorded the maximum occurrences of network failures. IT department and management are the next at risk departments as far as the rate of network downtimes are concerned whilst the other departments ensue as illustrated in Figure 8.

The engineering departments leads the rate of downtimes because of their closeness to the gold processing plant, proneness to vibrations due to the plant's operation and the movement of heavy mine machineries, LAN infrastructural relocations and geography.

For the finance and procurement departments, LAN infrastructural relocations and physical obstructions accounted for their frightening network downtime rate. Averagely, the minimum length of LAN network downtimes is four (4) hours in every two days, and 54% of the respondents become idle or redundant during this time.

### f) *Quantification of Actual Network Downtimes Losses*

On the basis of six working days in a week, the actual average weekly network downtime according to Table 2 is 14.31 hours per employee (4.77 hours in every two days). Quantifying the loss due to this network downtime for 900 employees (the minimum number proposed by the IT staff) is shown below.

An average monthly salary per employee proposed by finance departments is GHc1, 000.00. The hourly labour loss is:

$$\textit{Hourly labour loss due to network downtime} =$$

$$= \frac{1000}{30\,days \times 8\,hours} = GHc\;4.17$$

$$\textit{Total monthly labour loss incurred} = 4.17 \times 14.31 \times 4 \times 900$$

$$= GHc\;214{,}821.72$$

$$\textit{Total yearly labour loss incurred} = 12 \times 214{,}821.72 =$$

$$= GHc\;2{,}577{,}860.64$$

Note that, the estimated labour loss due to network downtimes of GHc2,577,860.64 excluded the cost of network reconstruction and stationeries due absence of the network. This loss is too high to neglect as a company, irrespective of their annual incomes.

The greatest want of the studied mining companies should be the want of in-house IT skills/experts who can effectively handle the specially-designed and mining-friendly new technologies with improved and robust LAN/WLAN network infrastructure.

## IV. Conclusion and Recommendation

### a) *Conclusion*

Over four productive hours in every two days per employee for more than 1000 employees (54%) are lost due to network failures/downtimes. This man-hour loss to talling GHc2, 577, 860.64 (USD 1,288,930.32) annually is mutely charged against productivity. Logically, it cannot be overemphasized that the amount contributes significantly to productivity loss irrespective of the company's annual profit. This affirms the fact that "increased LAN/WLAN network infrastructural relocations resulting in LAN/WLAN network downtimes in mining operational environment vis-à-vis some in-house obstacles decreases mine productivity". Mining operations are supported by software applications accessed through a network. Wired and Wireless media network connectivity enables effective communication in the mines. Thus, profitable mining operations depend on effective communication. When data network like LAN/WLAN shuts down or becomes unavailable, safety and productivity are compromised due to long employee productive hour loss. In the worst case, the entire operation must be suspended.

### b) *Recommendations*

Pragmatically, random LAN infrastructural relocations, obstructions to line-of–sights of wireless medium of communication, attenuation in wireless information signal strength due to atmospheric dust concentrations and vibrations from numerous sources are inevitable. Nonetheless, a better alternative must be considered.

- Against this background, the following recommendations are being made: in the short term;
- Well-planned preventive maintenance schedule for all existing LAN/WLAN infrastructure.
- Basic ICT Training modules introduced into the regular training modules for management and all employees.
- Technically qualified IT experts made part of management and business improvement department to handle pertinent IT projects and issues.
- Further research into the impact of vibration on LAN communication network infrastructure.

## V. Acknowledgement

EE and CKA Thanks the University of Mines and Technology, Tarkwa, for their support. Also, the authors acknowledge the support received from Goldfields Ghana Ltd, Tarkwa and Damang Gold Mines and Anglogold Ashanti Iduapriem Mine.

## References Références Referencias

1. Tanebaum, A. S. (2003).Computer Networks, Fourth Edition. Prentice Hall.
2. Collins, D., & Smith, C. (2001). 3G Wireless Networks, New York: McGraw-Hill.
3. Dajab, D. D (2006). Perspectives on the Effects of Harmattan on Radio Frequency Waves. J. Appl. Sci. Res., 2 (11): 10141018.
4. Dimari, G. A. Maitera, O. N. Waziri, M. &Hati, S. S. (2008). Pollution Synergy from Particulate Matter Sources: The Harmattan, Fugitive Dust and Combustion Emissions in Maiduguri Metropolis, Nigeria. European Journal of Scientific Research, ISSN 1450-216X Vol.23 No.3, pp.465-473.
5. Folaponmile A and Sani M. S. (2011). Empirical model for the prediction of mobile radio cellular signal attenuation in harmattan weather.
6. Breuning-Madsen, H. and T. W. Awadzi, (2005). Harmattan dust deposition and particle size in Ghana. Journal Catena (63: 1), pp 23-38.
7. Anderson, H., Hicks, T. & Kirtner, J. (2008). "The Application of Land Use/Land Cover (Clutter) Data to Wireless Communication System Design", EDX *Wireless, LLC, Eugene*, Oregon USA.
8. Ashish, S. & Prashant, J. (2010). "Effects of Rain on Radio Propagation in GSM". International Journal of Advanced Engineering & Applications, Delhi.
9. Shneiderman, B. & C. Plaisant (2010). Designing the User Interface: Strategies for Effective Human-Computer Interaction, Fifth Edition, addition Wesley Imprint.
10. Zikmund, W. G. (1994) Exploring Marketing Research, Fort Worth: Dryden Press.
11. Creswell, J. W. (1994). Research design: Qualitative & Quantitative Approaches, USA: Sage Publications.
12. Yin, R. K., (1994). Case Study Research: Design and Methods, applied Social Research Methods Series, 2nd Ed. Sage Publishing, Newbury Parl California.
13. Hair, J. F., Jr., Babin, B., Money, A. H., & Samouel, P. (2003). Essentials of business research methods. New York: Wiley.
14. Holme, I. M. & Solvang, B. K. (1991) Research Methods: Qualitative and Quantitative Methods, About Student: Lund.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Quantifying COTS Components Selection using Multi Criteria Decision Analysis Method PROMETHEE

By Kulbir Kaur & Harshpreet Singh

*Lovely Professional University, India*

*Abstract-* Component Based Development relies on already existing components to develop the system. It offers various advantages as increase in productivity, reduced development effort and time. The biggest challenge is to select the appropriate component from number of alternatives based on the quality parameters. In this paper COTS component selection is reduced to a multi criteria decision problem by quantifying it with PROMETHEE method. PROMETHEE is an outranking method which better supports the evaluation and selection from various alternatives based on the functional and non-functional requirements. The aim of this paper is to show the application of PROMETHEE in evaluating, analysing and selecting the appropriate COTS component with respect to requirements. The paper also discusses the procedure and benefits of using PROMETHEE method over the other MCDA methods.

*Keywords: COTS, CBD, MCDA, PROMETHEE, AHP, WSM.*

*GJCST-C Classification: K.6.3*

# Quantifying COTS Components Selection using Multi Criteria Decision Analysis Method-PROMETHEE

Kulbir Kaur α & Harshpreet Singh σ

*Abstract-* Component Based Development relies on already existing components to develop the system. It offers various advantages as increase in productivity, reduced development effort and time. The biggest challenge is to select the appropriate component from number of alternatives based on the quality parameters. In this paper COTS component selection is reduced to a multi criteria decision problem by quantifying it with PROMETHEE method. PROMETHEE is an outranking method which better supports the evaluation and selection from various alternatives based on the functional and non-functional requirements. The aim of this paper is to show the application of PROMETHEE in evaluating, analysing and selecting the appropriate COTS component with respect to requirements. The paper also discusses the procedure and benefits of using PROMETHEE method over the other MCDA methods.

*General Terms: selection, alternative, criteria, rank, degree, preference, profile.*

*Keywords: COTS, CBD, MCDA, PROMETHEE, AHP, WSM.*

## I. Introduction

Component Based Development (CBD) relies on reusable COTS components to build the software systems. Before integrating the components into the system, the components should be quantified according to the non-functional and functional requirements.

With the rapid growing and changing of technology, number of products or tools entering in the market also increases. So it becomes a big challenge to select the best component from a number of alternative components and to build a trust on the selected components.

Component selection and evaluation is a multi criteria problem in which a component from various alternatives is to be selected which best satisfies the maximum criteria than others. A chosen option should have greater rank on all criteria than others.

Various methods can be used as a solution of this problem like OSTO [2], CARE [8], AHP [3], WSM [3], Utility Theory [1], SMART [1], DesCOTS [9], UnHOS [7] etc. Multi Criteria Decision Analysis methods help the decision maker to select the best option from number of multi criteria alternatives which best scores on multiple criteria. PROMETHEE is a multi criteria method proposed by JP Brans in 1982 [6]. It can be applied for the analysis and selection of components and solutions in various kinds of fields like Banking, Industrial Location, Manpower planning, Water resources, Investments, Medicine, Chemistry, Health care, Tourism, Ethics in OR, Dynamic management [6]. It can be applied to selection and evaluation of COTS components while making the decision to select components from repository to develop the software system. The aim of this paper is to apply PROMETHEE on the selection and evaluation of software packages and its benefits over others multi criteria methods.

## II. Literature Review

COTS-Aware Requirements Engineering and Software Architecting (CARE/SA) proposed by Lawrence [8] for evaluating, matching and selecting of COTS components. CARE/SA method uses the architectural aspects, functional aspects and non-functional aspects of COTS components. It indicates that each component is represented by the unique attributes which consists of its architectural, functional and non-functional aspects.

Hamdy Ibrahim et al. in [7] proposed a method named 'UnHOS' (Uncertainty Handling in COTS Selection) method for the evaluation of COTS components and takes into account their uncertainty. It uses Analytic Hierarchy Process (AHP) for the evaluation of COTS components and Bayesian Belief Network (BBN) to indicate their uncertainty. It also presents a tool to support the usability of the UnHOS method.

Anil Jadhav et al. in [3] tells that Multi Criteria Decision Making Methods helps the decision makers to solve the problem of selection and evaluation of software components in which problem is defined as a collection of multiple criteria that needs to be taken into account. It gives the overview of Multi Criteria Decision Making Methods like: Analytic Hierarchy Process (AHP), Weighted Scoring Method (WSM) and Hybrid Knowledge Based System (HKBS). It compares the three approaches and concludes that HKBS is better than AHP and WSM.

*Author α: Student, School of CSE and IT, Lovely Professional University, Phagwara, Punjab. e-mail: gill.kulbir11@gmail.com*

*Author σ: Assistant Professor, School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab. e-mail: harsh08walia@gmail.com*

Arvinder Kaur et al. in [2] provide a brief overview of the evolutionary techniques. It also derives a hierarchical decomposition method to draw goals from that impact factors. It introduces OSTO method for the selection of software components which compares the scores and cost associated to each alternative and their relative comparison. It introduces various factors in the selection of reusable software components. It also presents the evaluation criteria based on various classifications as functional requirements, product quality attributes, strategic concerns and architecture and domain compatibility. It gives the result of two case studies using OSTO method. The component which have good quality assurance score is selected for consideration.

## III. Multi Criteria Decision Analysis Method

Multi criteria problem involves the selection of the best option from a number of alternatives on the basis of multiple criteria satisfaction with higher degree. As component selection is a multi-criteria problem, there are number of alternatives for the solution of problem and we have to select a candidate component which best suits for the solution on the basis of satisfying maximum criteria than others with higher degree. So problem can be formulated as:

$\max \{c_1an, c_2an……ckan|a_n \in A\}$.

Let $A= \{a_1, a_2, a_3……………….a_n\}$ be the set of 'n' alternatives for the solution of the problem.

$C= \{c_1, c_2, c_3…………………c_k\}$ be the set of 'k' criteria as a basis of evaluation and selection.

Let $w_1,w_2,w_3………w_k$ be the weight of each criterion respectively.

Each multi criteria decision analysis method proceeds with the decision table. Decision Table is shown in Table 1. Each column denotes the criteria, each row denotes the alternatives and 'ckan' represents the score of alternative 'n' on criterion 'k'.

*Table 1 :* The decision table

| Alternative | Criteria | | | |
|---|---|---|---|---|
| | $w_1$ | $w_2$ | …. | $w_k$ |
| | $c_1$ | $c_2$ | …. | $c_k$ |
| $a_1$ | $c_1a_1$ | $c_2a_1$ | …. | $cka_1$ |
| $a_2$ | $c_1a_2$ | $c_2a_2$ | …. | $c_ka_2$ |
| . | . | . | …. | . |
| . | . | . | …. | . |
| $a_n$ | $c_1a_n$ | $c_2a_n$ | …. | $c_ka_n$ |

### a) *PROMETHEE Method*

There is need to have a method which is simpler and better helps in decision making while obtaining the solution of multi objective selection of trusted components from the number of available alternatives. As COTS components selection is a multi-criteria problem. PROMPTHEE solves the problem in an optimal way with additional benefits than other MCDA methods.

PROMETHEE is Preference Ranking Organisation Method for Enrichment Evaluation. PROMETHEE is a multi criteria decision analysis method. It is an outranking method based on pair wise comparison of alternatives. It was developed by JP Brans in 1982[6]. Originally it was developed as PROMETHEE-1 (partial ranking) and PROMETHEE-2 (complete ranking).Later PROMETHEE-3 (ranking based on intervals) and PROMETHEE-4 (continuous case) were developed. PROMETHEE-5 (MCDA includes segmentation constraints) and PROMETHEE-6 (represents human brain) are also there. PROMETHEE is based on mathematical properties [6]. It can be applied on various fields for the selection and evaluation of winning solution in a multi criteria problem.

Steps for solving multi criteria problem with this method is as follows:

1. Determination of available alternatives to solve the problem.

Let $A= \{a_1, a_2, a_3……………….a_n\}$ be the set of 'n' alternatives for the solution of the problem.

2. Determination of evaluation criteria. Let $C= \{c_1, c_2, c_3…………………c_k\}$ be the set of 'k' criteria as a basis of evaluation and selection.
3. Problem statement stated as $\max\{c_1(an),c_2(an),c_3(an),………c_j(an)…….c_k(a_n)|a_n \in A\}$ Where '$c_k(a_n)$' represents the value of alternative '$a_n$' on the criterion 'k'.
4. Create an evaluation table or (n*k) matrix with 'n' rows (number of alternatives) and 'k' columns (number of evaluation criteria) and place the score value of each alternative based on each criterion i.e. '$c_ka_n$'.
5. Assign weight to each criterion i.e. $w_j$ where $(j=1,2,3…..k)$ and $w_1+w_2+w_3…..w_k=1$.
6. Find the difference between each pair of alternatives based on each criterion i.e. $d_j(a,b)= c_j(a)-c_j(b)$ where $(j=1, 2…k)$ and $(a,b \in A)$.
7. Find the preference of the one alternative over the other as a function of difference between each pair of alternatives based on each criterion i.e. $Pj (a,b)= F_j [d_j(a,b)]$ where $(a,b \in A)$ and $(j=1, 2…k)$ and $0 \le P_j(a,b) \le 1$. In case of minimizing the criteria preference $P_j(a,b)=F_j[-d_j(a,b)]$. Preference function values can be taken on the basis of particular criterion function and the parameter value which you have selected.
8. Calculate the degree to which preferred option is better than other alternative on all criteria i.e. $\pi(a,b)= P_1 (a,b)w_1+P_2(a,b)w_2+………P_k(a,b)w_k$. And $0 \le \pi (a,b) \le 1$.
9. If the degree of preference nearly equals to 'zero'; it means there is weak preference of alternative 'a'

over alternative ‘b’. And if the degree of preference nearly equals to ‘one’; it means there is strong preference of alternative ‘a’ over alternative ‘b’.

10. Calculate the positive and negative outranking flow of each option and then compute the net outranking flow of the option and if it comes out to be greater than ‘zero’ then the option outranking the other options and if lower than ‘zero’ then it means that the option is outranked by the other options on all criteria.

    Positive outranking flow (option outranks others):

    $\Phi+(a)=1/(n-1)[\pi(a1,a2)+\pi(a1,a3)\ldots\ldots\pi(a1,an)]$

    Negative outranks flow (option is outranked by others):

    $\Phi-(a)=1/(n-1)[\pi(a2,a1)+\pi(a3,a1)\ldots\ldots\pi(an,a1)]$

    Net outrank flow of an option:

    $\Phi(a)= \Phi+(a)- \Phi-(a)$.

    We can say that ‘a’ is preferred over ‘b’ if $\Phi(a)> \Phi(b)$ and $0\leq\Phi(a) \leq1$. Moreover$\{\Phi(a1)+ \Phi(a2)+\ldots.. \Phi(an)=0\}$

11. Obtain the outrank flow of each option on each criterion as:

    $\Phi j(a)=1/(n-1)[(P1(a,b) - P1(b,a))+(P2(a,b)-P2(b,a)) +P3(a,b)-P3(b,a))\ldots\ldots+ Pk(a,b)-Pk(b,a))]$.

12. Obtain the profile of an alternative on all the criteria as:

    $\Phi(a)= \Phi1(a)w1+ \Phi2(a)w2+\ldots. \Phi k(a)wk$.

Profile of an alternative indicates the quality of an alternative on each criterion. Profile is shown in figure 1.

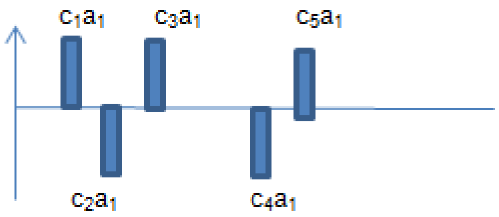


*Figure 1 :* Profile of an alternative

13. Select the alternative which has high ‘Φ(a)’. Values of ‘Φ(a)’ gives the complete rank of the alternatives.

### *b) Promethee on Cots Components Selection*

While developing system from COTS components it becomes very difficult to select best one if number of alternatives are available and to evaluate those alternatives. Application of PROMETHEE methodology on the COTS components selection better supports us in decision making.

Suppose a set of software components i.e. Alternatives set as A={A1,A2,A3,A4,A5} and evaluation criteria set as C={Performance, Reliability, Maintainability, Cost, Integrability} and the weight of each criterion respectively is as:0.3,0.2,0.1,0.2,0.2.

Let criteria can be written as C1, C2, C3, C4 and C5.

Scale and units for criteria are as follows:

C1: VG, G, A, B, VB
C2:no. of failures per 1000 hours of service
C3: VG, G, A, B, VB
C4: Rs. (Rupees)
C5: VG, G, A, B, VB

Where VG, G, A, B, VB stands for very good, good, average, bad, very bad.

Score for each grade is as in table 2.

*Table 2 :* Grade scores

| Grade | VG | G | A | B | VB |
|---|---|---|---|---|---|
| Score | 5 | 4 | 3 | 2 | 1 |

Evaluation table is shown in table 3.

*Table 3 :* Evaluation table

| | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| | 0.3 | 0.2 | 0.1 | 0.2 | 0.2 |
| A1 | 3 | 2 | 4 | 1000 | 5 |
| A2 | 2 | 1 | 3 | 1200 | 4 |
| A3 | 5 | 3 | 5 | 900 | 3 |
| A4 | 4 | 7 | 3 | 1000 | 2 |
| A5 | 4 | 2 | 2 | 1100 | 1 |

Preference function may be used as Usual, U-Shaped, V-Shaped, Level, Gaussian or V-Shape with indifference criterion function. Let in the example U-Shaped criterion function is taken for C2 and C4. Level criterion function is taken for C1, C3 and C5. Level and U-Shaped criteria are shown in figure 2 and figure 3 respectively.

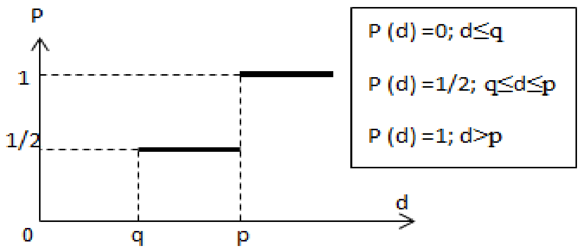


*Figure 2 :* Level Criterion

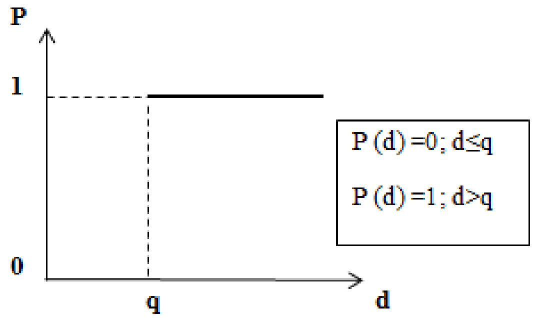


*Figure 3 :* U-Shaped Criterion

Let parameter values for each criterion is as follows:
For C1, C3 and C5; q=2 and p=4
For C2; q=4

For C4; q=100

Relative difference between alternatives on each criterion is shown in tables 4, 5, 6, 7 and 8.

*Table 4 :* Difference between alternatives with respect to performance

| $d_1(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | 1 | -2 | -1 | -1 |
| A2 | -1 | 0 | -3 | -2 | -2 |
| A3 | 2 | 3 | 0 | 1 | 1 |
| A4 | 1 | 2 | -1 | 0 | 0 |
| A5 | 1 | 2 | -1 | 0 | 0 |

*Table 5 :* Difference between alternatives with respect to reliability

| $d_2(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | -1 | 1 | 5 | 0 |
| A2 | 1 | 0 | 2 | 6 | 1 |
| A3 | -1 | -2 | 0 | 4 | -1 |
| A4 | -5 | - 6 | -4 | 0 | -5 |
| A5 | 0 | -1 | 1 | 5 | 0 |

*Table 6 :* Difference between alternatives with respect to maintainability

| $d_3(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | 1 | -1 | 1 | 2 |
| A2 | -1 | 0 | -2 | 0 | 1 |
| A3 | 1 | 2 | 0 | 2 | 3 |
| A4 | -1 | 0 | -2 | 0 | 1 |
| A5 | -2 | -1 | -3 | -1 | 0 |

*Table 7 :* Difference between alternatives with respect to cost

| $d_4(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | 200 | -100 | 0 | 100 |
| A2 | -200 | 0 | -300 | -200 | -100 |
| A3 | 100 | 300 | 0 | 100 | 200 |
| A4 | 0 | 200 | -100 | 0 | 100 |
| A5 | -100 | 100 | -200 | -100 | 0 |

*Table 8 :* Difference between alternatives with respect to integrability

| $d_5(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | 1 | 2 | 3 | 4 |
| A2 | -1 | 0 | 1 | 2 | 3 |
| A3 | -2 | -1 | 0 | 1 | 2 |
| A4 | -3 | -2 | -1 | 0 | 1 |
| A5 | -4 | -3 | -2 | -1 | 0 |

Preference function value of each alternative over other on all criteria is shown in table 9, 10, 11, 12 and 13.

*Table 9 :* Preference value on performance

| $P_1(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | 0 | 0 | 0 | 0 |
| A2 | 0 | 0 | 0 | 0 | 0 |
| A3 | .5 | .5 | 0 | 0 | 0 |
| A4 | 0 | .5 | 0 | 0 | 0 |
| A5 | 0 | .5 | 0 | 0 | 0 |

*Table 10 :* Preference value on reliability

| $P_2(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | 0 | 0 | 1 | 0 |
| A2 | 0 | 0 | 0 | 1 | 0 |
| A3 | 0 | 0 | 0 | 0 | .5 |
| A4 | 0 | 0 | 0 | 0 | 0 |
| A5 | 0 | 0 | 0 | 1 | 0 |

*Table 11 :* Preference value on maintainability

| $P_3(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | 0 | 0 | 0 | .5 |
| A2 | 0 | 0 | 0 | 0 | 0 |
| A3 | 0 | .5 | 0 | .5 | .5 |
| A4 | 0 | 0 | 0 | 0 | 0 |
| A5 | 0 | 0 | 0 | 0 | 0 |

*Table 12 :* Preference value on cost

| $P_4(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | 1 | 0 | 0 | 0 |
| A2 | 0 | 0 | 0 | 0 | 0 |
| A3 | 0 | 1 | 0 | 0 | 1 |
| A4 | 0 | 1 | 0 | 0 | 0 |
| A5 | 0 | 0 | 0 | 0 | 0 |

*Table 13 :* Preference value on integrability

| $P_5(a,b)$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | 0 | .5 | .5 | .5 |
| A2 | 0 | 0 | 0 | .5 | .5 |
| A3 | 0 | 0 | 0 | 0 | .5 |
| A4 | 0 | 0 | 0 | 0 | 0 |
| A5 | 0 | 0 | 0 | 0 | 0 |

Degree of preference of one alternative over other is shown in table 14.

*Table 14 :* Degree of preference Π(a,b)

| $\Pi_{(a,b)}$ | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | 0 | .20 | .10 | .30 | .15 |
| A2 | 0 | 0 | 0 | .30 | .10 |
| A3 | .15 | .40 | 0 | .05 | .35 |
| A4 | 0 | .35 | 0 | 0 | 0 |
| A5 | 0 | .15 | 0 | .20 | 0 |

Positive, negative and net outrank flow of each alternative is shown in table 15.

*Table 15 :* Positive, negative and net outrank flow

| | $\Phi^+(a)$ | $\Phi^-(a)$ | $\Phi(a)$ |
|---|---|---|---|
| A1 | .1875 | .0375 | .1500 |
| A2 | .1000 | .2750 | -0.1750 |
| A3 | .2375 | .0250 | .2125 |
| A4 | .0875 | .2125 | -0.1250 |
| A5 | .0875 | .1500 | -0.0625 |

PROMETHEE-1 Partial ranking of each alternative is shown in figure 4 and PROMETHEE-2 Complete ranking of each alternative is shown in figure 5.

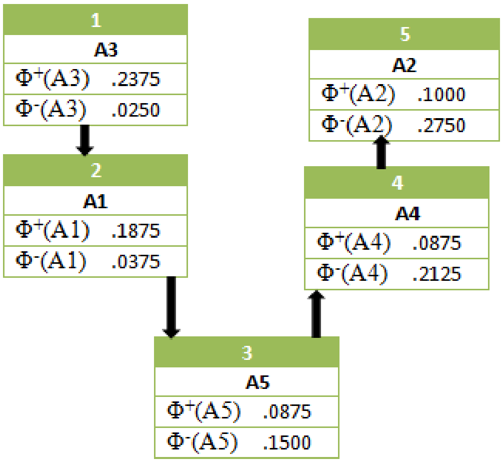


*Figure 4 :* Partial ranking

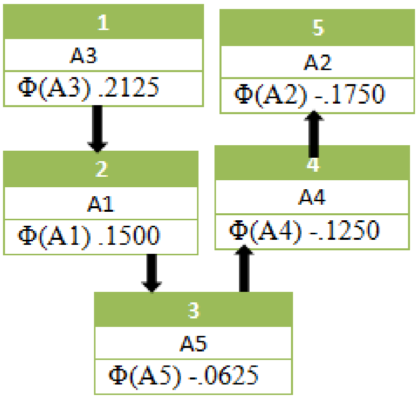


*Figure 5 :* Complete ranking

Profile of each alternative on all criteria is shown in table 16.

*Table 16 :* Profile of alternative

| | $\Phi_1(a)$ | $\Phi_2(a)$ | $\Phi_3(a)$ | $\Phi_4(a)$ | $\Phi_5(a)$ |
|---|---|---|---|---|---|
| A1 | -.125 | .250 | .125 | .250 | .375 |
| A2 | -.375 | .250 | -.125 | -.750 | .250 |
| A3 | .250 | 0 | .375 | .500 | 0 |
| A4 | .125 | -.750 | -.125 | .250 | -.250 |
| A5 | .125 | .250 | -.250 | -.250 | -.375 |

Profile of alternative A1 on all criteria is shown in figure 6.

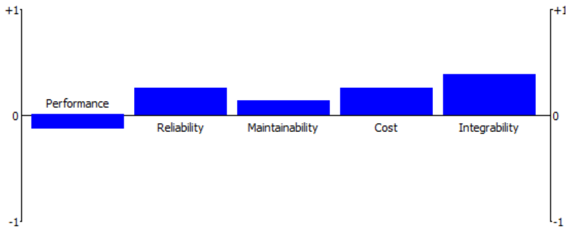


*Figure 6 :* Profile of A1

Profile of alternative A2 on all criteria is shown in figure 7.

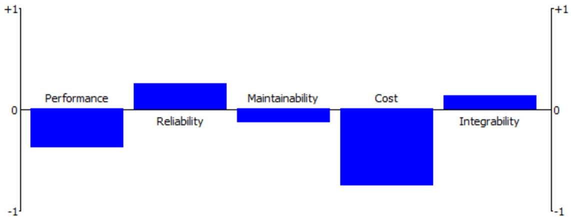


*Figure 7 :* Profile of A2

Profile of alternative A3 on all criteria is shown in figure 8.

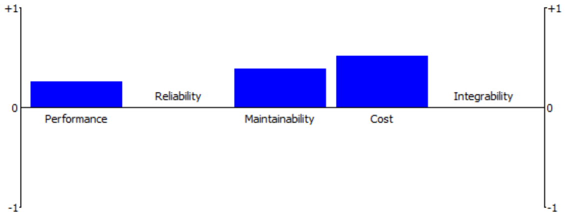


*Figure 8 :* Profile of A3

Profile of alternative A4 on all criteria is shown in figure 9.

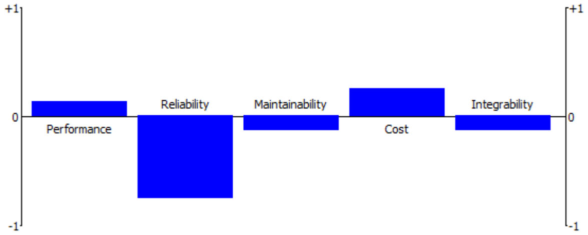


*Figure 9 :* Profile of A4

Profile of alternative A5 on all criteria is shown in figure 10.

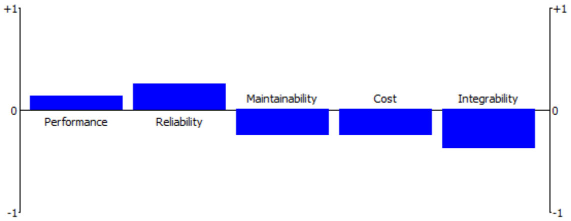


*Figure 10 :* Profile of A5

Ranking of all alternatives on all criteria is shown in figure 11.

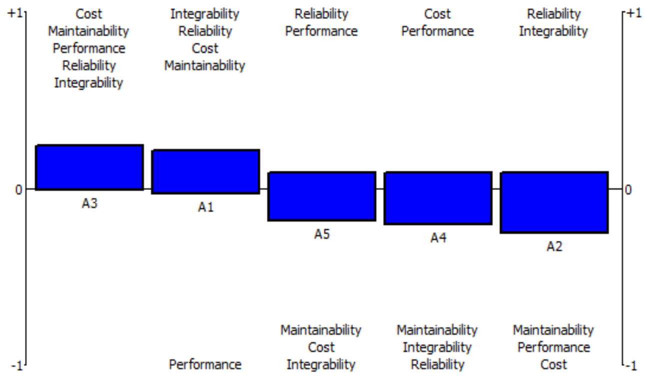


*Figure 11 :* Ranking of all alternatives on all criteria

### *c) Benefits of using Promethee*

PROMETHEE methodology helps in decision making better than other MCDA methods in number of ways:

1. The supporting software packages of PROMETHEE like D-Sight, PROMCALC, Decision Lab, Visual PROMETHEE etc. are very user friendly.
2. PROMETHEE-GDSS supports group decision making in this the final solution is obtained by weighted sum of net outrank flow of each alternative.
3. PROMETHEE provides partial, complete, interval based ranking of alternatives.
4. Unlike other MCDA methods, PROMETHEE's preference degree tells us the degree by which an option is preferred over other.
5. PROMETHEE needs very less input for further operations as compare to other MCDA methods.
6. Unlike other MCDA methods, new alternatives can be added without doing much change in others.
7. PROMETHEE does not include normalization for normalizing the units of measurement of each criterion so there are fewer chances of errors as compare to many other MCDA methods.
8. PROMETHEE's extensions can be used as sorting purposes.

## IV. Conclusion

Component selection is a wide comparison of components using a common set of criteria. Selecting the appropriate and relevant component significantly reduces the chances of risks associated with the COTS components with no source code available with them and improves the corporate competitiveness. Using PROMETHEE-GAIA methodology for the complete ranking of alternatives help decision makers to choose and analyse the highest rank component on all criteria and help to build confidence on the selected component.

## References Références Referencias

1. Hutchinson, J., & Kotonya, G. (2006, August). A Review of Negotiation Techniques in Component Based Software Engineering. In Software Engineering and Advanced Applications, 2006. SEAA'06. 32nd EUROMICRO Conference on(pp. 152-159). IEEE.
2. Kaur, A., & Mann, K. S. (2010). Component Selection for Component-Based Software Engineering. International Journal of Computer Applications, 2(1), 109-114.
3. Jadhav, A., & Sonar, R. (2009, December). Analytic Hierarchy Process (AHP), Weighted Scoring Method (WSM), and Hybrid Knowledge Based System (HKBS) for Software Selection: A Comparative Study. In Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on (pp. 991-997). IEEE.
4. Becker, C., Kraxner, M., Plangg, M., & Rauber, A. (2013, January). Improving decision support for software component selection through systematic cross-referencing and analysis of multiple decision criteria. In System Sciences (HICSS), 2013 46th Hawaii International Conference on (pp. 1193-1202). IEEE.
5. Halim, A., Sudrajat, A., Sunandar, A., Arthana, I. K. R., Megawan, S., & Mursanto, P. (2011, December). Analytical Hierarchy Process and PROMETHEE application in measuring object oriented software quality. InAdvanced Computer Science and Information System (ICACSIS), 2011 International Conference on (pp. 165-170). IEEE.
6. Brans, J. P., & Mareschal, B. (2005). PROMETHEE methods. In Multiple criteria decision analysis: state of the art surveys (pp. 163-186). Springer New York.
7. Ibrahim, H., Elamy, A. H. H., Far, B. H., & Eberlein, A. (2011). UnHOS: A Method for Uncertainty Handling in Commercial Off-The-Shelf (COTS) Selection.International Journal of Energy, Information & Communications, 2(3).
8. Chung, L., Cooper, K., & Courtney, S. (2004). COTS-Aware Requirements Engineering and Software Architecting. In Software Engineering Research and Practice (pp. 57-63).
9. Grau, G., Carvallo, J. P., Franch, X., & Quer, C. (2004, August). DesCOTS: a software system for selecting COTS components. In Euromicro Conference, 2004. Proceedings. 30th (pp. 118-126). IEEE.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Evaluation the Quality of Software Design by Call Graph based Metrics

By Sanjeev Kumar Punia, Dr. Anuj Kumar & Amit Sharma

*IIMT College of Engineering, India*

*Abstract-* The prediction of software defects was introduced to support development and maintenance activities to improve the software quality by finding errors early in the software development. It facilitates maintenance in terms of effort, time and more importantly the cost prediction for software evolution and maintenance activities.

In this paper, we evaluate the quality related attributes in developed software products. The software call graph model is also used for several applications in order to represent and reflect the degree of their complexity in terms of understandability, testability and maintainability efforts. The extracted metrics are investigated for the evaluated applications in correlation with bugs collected from customers bug reports. Those software related bugs are compiled into datasets files to use as an input to a data miner for classification, prediction and association analysis.

Finally, the analysis results is evaluated in terms of finding the correlation between software products bugs and call graph based metrics. We find that call graph based metrics are appropriate to detect and predict software defects so that the activities of testing and maintenance stages become easier to estimate or assess after the product delivery.

*Keywords: software testing, software metrics, coupling metrics, call graph based metrics, defects prediction and software maintainability.*

*GJCST-C Classification: K.6.3*

# Evaluation the Quality of Software Design by Call Graph based Metrics

Sanjeev Kumar Punia [α], Dr. Anuj Kumar [σ] & Amit Sharma [ρ]

*Abstract-* The prediction of software defects was introduced to support development and maintenance activities to improve the software quality by finding errors early in the software development. It facilitates maintenance in terms of effort, time and more importantly the cost prediction for software evolution and maintenance activities.

In this paper, we evaluate the quality related attributes in developed software products. The software call graph model is also used for several applications in order to represent and reflect the degree of their complexity in terms of understandability, testability and maintainability efforts. The extracted metrics are investigated for the evaluated applications in correlation with bugs collected from customers bug reports. Those software related bugs are compiled into datasets files to use as an input to a data miner for classification, prediction and association analysis.

Finally, the analysis results is evaluated in terms of finding the correlation between software products bugs and call graph based metrics. We find that call graph based metrics are appropriate to detect and predict software defects so that the activities of testing and maintenance stages become easier to estimate or assess after the product delivery.

*Keywords: software testing, software metrics, coupling metrics, call graph based metrics, defects prediction and software maintainability.*

## I. Introduction

The human abilities and creativities play a significant role in producing and directing the software products in software development life cycle with the help of the tools and methodologies. However, humans are also the main source of the errors and defects that occur in the software and discovered before or after the delivery of the product. The production of defect free software and projects is impossible. However, software developers struggle to keep such possible defects at minimum. Finding and fixing the defects and errors after delivery usually cost a large amount of the project budget and resources specially if compared with detecting them earlier. As try to predict the defects early is valuable specially if detected before the delivery of the software where that can also help the project to success in terms of cost and quality.

The coupling metrics play an important role in many software development and maintenance activities such as effort estimation, improving the quality of the software products, test planning and reducing future maintenance. These metrics assess the software quality by supporting the quality related factors after evaluating error proneness, changeability and reusability. The most relevant tools are available as independent or the part of a development environment to compute the coupling metrics statically by tracing possible problems in the source code.

The call graphs metrics represent the relationship between the modules and reflect the degree of complexity of the software. It also helps to find some software metrics such as coupling and cohesion metrics. In general, one way to reduce cost through defects prediction is by using software metrics. Particularly the call graph based metrics is used to predict and improve possible problems in software design and in coding finally.

In this research, we tried to evaluate the effectiveness and power of call graph based metrics in prediction and detection the defects in software products. A tool is developed to generate call graph attributes and metrics by using open source projects. We select three applications as J Edit 4.2, Velocity 1.4 and Velocity 1.6 based on two factors (i) open source projects and (ii) these projects include users bug reports for actual evaluation of the software products. This paper include, the programmed and evaluated call graph based metrics as LOC, Fan In, Fan Out, SGBR and IFC. The LOC, Fan In and Fan Out metrics are known and popular while SGBR and IFC not so popular but after that also we implement in our tool.

This paper is organized into six sections as follows: Section 2 introduces topic and research related studies. Section 3 describes the methodology steps. Section 4 presents the adopted analysis and evaluation measurements. Section 5 shows the conducted results of the experiments and finally Section 6 presents the conclusion and inference from the paper.

## II. Literature Review

Many empirical studies used call graph based model for developing the derive dependency metrics especially code and size metrics. Multiple authors/researchers proposed different ways to utilize

*Author α: IIMT College of Engineering, Greater Noida.*
*e-mail: puniyasanjeev@hotmail.com*
*Author σ: Accurate Institute of Engineering & Technology, Greater Noida. e-mail: Anujkumar74@rediffmail.com*
*Author ρ: IIMT College of Engineering, Greater Noida.*
*e-mail: Amit.krsharma123@gmail.com*

call graph based dependency metrics to improve the software quality by providing information for defect prediction and estimation. We list some related work in each step that has taken in our project and developed tool in the following sections.

### *a) Call Graph Model*

Many researchers studied software modeling and found that modeling techniques are grouped into broadly two categories as (i) graphical modeling techniques and (ii) textual modeling techniques. Graphical modeling technique use a diagram with named symbols that represent the components, the symbols connecting arcs represent the relationship and other notations to represent the constrains. Textual modeling technique use standardized notations and keywords to define major aspects of software product call graph.

J. Dollner and Bohnet *et.al.* [1] Considered the extracting of process call dependencies as one of the most important step in the reengineering process. Therefore they built a tool based on OINK framework for call graph extraction. In addition, the tool also provides a set of hierarchal data, call type information methods definitions and output this information to impor Table formatted file.

D. Reniers *et.al.* [2] Made an enhancement in hierarchical edge bundling (HEB) technique and named candidate visualization (CV) technique in their framework. So they build an experiment to compare their enhancement hierarchical edge bundling and tulip graph visualization framework with several large systems like Bison, Mozilla Firefox, OINK and conclude that hierarchical edge bundling scheme perform better in typical comprehension tasks.

M. Jahromi and E. Honar *et.al.* [3] introduced a new framework for call graph construction for program analysis. They choose ASM and soot a byte code reader for their environment to store information about the structure of the codes such as classes, methods, files and statements.

They also proposed a framework where three algorithms have been implemented for call graph construction i.e. CHA, RTA and CTA and finally they conclude by an experimental study on a verity of source code programs by comparing two byte code reader.

### *b) Code Metrics Extraction*

By analyzing both the source code of any software and extracting code metrics is considered as the main preprocess for the reengineering operation. This information provides a clear view about the complexities and difficulties of the software as well as divides the milestones tasks into phases in order to start the reengineering process easily. On the other hand many researchers considered the code metrics and the system complexity information as a good defect tracker. They setup a number of hypothesis related to defect probability and code metrics to prove the correlation between them but the hottest topics in this research is to define the set of metrics that we can considered them as the optimal defect predictor. The researchers shows many studies to define this set of metrics and try to view it's set as the perfect one that give justifications for their results. It is also find that code metrics, which plays a major role in many research fields and many tools deployed to handle extraction using different approaches.

F. Abreu and Baroni *et.al.* [4] Presented a new framework for metrics extraction by modeling the extraction data using UML Meta model called FLAME. They briefly mentioned the main characteristics of FLAME for fact extraction and recommended to use in firing a new tool for metrics extraction. The authors introduce an approach to formalize the metrics design in the optimal way where FLAME functions are used to extract well known sets of metrics as MOOD, MOOD2, MOOSE, EMOOSE and QMOOD metrics.

### *c) Defect Prediction from Source Code Metrics and System History*

A number of approaches have been deployed for defect prediction based on different criteria and information. Some researchers turn to find bugs in software code by analyzing software source code and compute its complexity. They extracted call graph based metrics from source code and used to decide which part or module of the software code likely to be defected. While other researchers prefer to use the system history to decide which part of them has a big defect probability especially when the application has many releases. They find that the system history is more accurate to predict defected parts of the system more than the code complexity extraction predict. While some studies support the two approaches together and use both in finding systems bugs.

A. Bernstein *et.al.* [5] compare the influence of different metrics used in defect prediction and defect prediction densities by using decision tree learners. They collected the needed data that is source code metrics and bugs report in the experiment from seven versions of open source code for Mozilla application. They applied J48 algorithm in WEKA data miner on the data set and setup a number of experiments to test their purposed hypothesis on defect predicting in software parts. They conclude that a simple tree learner can produce good results with various sets of input data.

N. Nagappan and T. Ball *et.al.* [6] introduce a new technique for prediction defect density by using code churn measures. The idea was drawn with a hypothesis that if code changes many times in the prerelease version then it also has a big chance to be defected in the post release. The authors build an experiment on W2K3 release with its service back and showed with its result that code churn is a good defect predictor. As they noticed that the increase of the code churn measures leads to an increase on the defect density in any software so they conclude that their metrics suit with line of code churned, deleted line of code, files churned, churn count and weeks of churn.

The aim of software developers is to evaluate the cost and quality of software before deliver to customers so that they can predict and find bugs or defects especially for critical systems.

## III. Methodology

Our methodology consists of six main phases as shown in Figure 1.

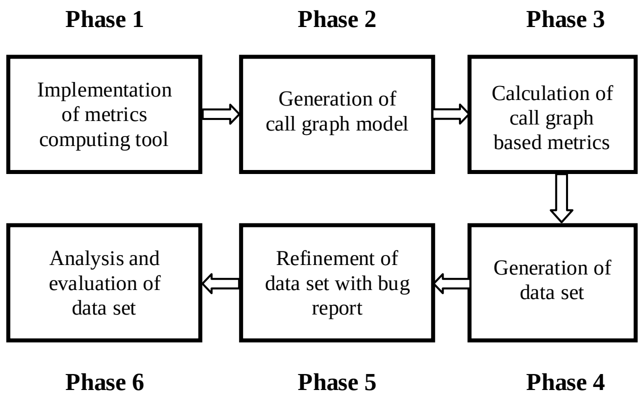

Phase 1
Phase 2
Phase 3
Implementation of metrics computing tool
Generation of call graph model
Calculation of call graph based metrics
Analysis and evaluation of data set
Refinement of data set with bug report
Generation of data set
Phase 6
Phase 5
Phase 4

*Figure 1 :* Methodology Diagram

This begin by phase 1 begin by "implementation of metric computing tool" to built a tool that can read source code of an application to compute some metrics coupling measurements. Phase 2 is "generation of call graph model" that utilizes the application model. Phase 3 is "calculation of call graph based metrics" used to compute some call graph based metrics for our application model. Phase 4 is "generation of data set" used to prepare data set consisted from metrics values for each class in application. Phase 5 is "refinement of data set with bug report" that assigns each data set record with its number of bugs. Finally, phase 6 is "*Analysis and Evaluation of Data Set using WEKA*" used for the purpose of evaluating its quality and find the correlation between its bug and call graph based metrics.

As the tool focus on extract the call graph based metrics so the developed tool generated data set does not contain bug attribute for each class. This phase is responsible to make some refinement on the comma separated values comma separate value (CSV) file.

Firstly, the tool automatically fill the bug attribute filed for each class by providing its previous comma separate value file for the same application under investigation and contains the bug report for each class. Then by mapping the name of classes between our comma separate value file and the previous worked comma separate value file. The source of previous comma separate value file gained from promise data repository which contains several data sets in comma separated values or attribute relation file (ARF) format. These files are created and prepared by researchers those worked at the topic of software defects prediction. In our research, we use bug attribute for the files which relate to the applications in our experiments.

## IV. Analysis and Evaluation

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper.

After refine the generated comma separate value file that represent the data set of our research with bug attribute then it is ready to analyze and evaluate using WEKA 3.7.5 tool as data miner. Here we apply J48, logistic model trees (LMT) and support vector machine (SMO) classifier algorithms. The decision tree algorithms are chosen as we want to look at classifiers that were easy to understand and validate the correlation between call graph based metrics and bugs.

### a) *Evaluation Measures*

The evaluation process of our testing tool depends on five matrices in term of call graph based metrics measurement. The five matrices are line of code (LOC), Fan In, Fan Out, call graph based ranking (CGBR) and information flow complexity (IFC) as shown in Table 1.

*Table 1 :* Measurement of call graph based metrics

| Metric Type | Measurement of Metrics |
|---|---|
| LOC | No of execu Table and non-commented lines of code for each function |
| Fan In | No of calling function list |
| Fan Out | No of called function list |
| CGBR | $(1\text{-}d) + d * \sum i\ CGBR(T_i)\ C(T_i)$ |
| IFC | IFC(M)=LOC(M)+ [Fan In(M)*Fan Out(M)]2 |

The metrics value for each type (LOC, Fan in, Fan Out, CGBR and IRC) depends on the functions that extracted from the application under investigation by which the higher metric value type achieves a higher complexity value. The values of metrics related to class level are computed by find the summation of all corresponded metrics to function level. For example: if we have 10 included functions at such class and each function has Fan In metrics value equal 1 then the class has Fan In metrics value equal the summation of all Fan

In metrics values related to functions of the class which equal to 10.

The five metrics we use in this research are related to size of the software or coupling and dependency between the components and functions of the application under investigation. LOC metric value represents the number of execu Table and non-commented lines of code. FanIn metric value for such function represents the number of function calling for a given function. Fan Out metric value for such function represents the number of function being called by a given function. CGBR metric depends on the page ranking algorithm that used by almost all the search engines where the ranking methodology is adopted to functions of the software.

This metric hypothesis that more frequently used functions and less frequently used modules should have different defects and bugs characteristic. In the equation used to compute CGBR value, value of d represents damping factor and refer to probability of such function being called or used and can be computed as the ratio of actual function calls to all possible function calls. CGBR (Ti) is the call graph based rank of module Ti which Call for given function. C(Ti) is the number of outbound calls of function Ti. IFC metric represents the measurement of the total level of information flow for a given function. The value of this type of metric depends on the values of metrics LOC, Fan In and Fan Out for the given function.

### b) *Principle Component Analysis using SPSS*

The purpose of this analysis is to show the correlate metrics in developed tool. The PCA analysis for call graph based metrics in developed tool results in 2 orthogonal dimension components were identified from 5 call graph based metrics that have Eigen value more than 1 as shown in Table 2. The variance of Eigen values data set explained by the PC in percent and the cumulative variance are provided for each PC where values above 0.6 are set in boldface. The 2 PCs capture 89.963% of the variance in the data set.

*Table 2 :* Rotated component matrix for developed tool

| | Component | |
|---|---|---|
| | 1 | 2 |
| Eigenvalue | 3.475 | 1.063 |
| % of Variance | 69.768 | 18.672 |
| Cumulative % | 68.362 | 89.235 |
| CGBR | 0.923 | - 0.112 |
| LOC | 0.905 | -0.131 |
| IFC | -0.036 | 0.923 |
| FanIn | 0.836 | 0.132 |
| FanOut | 0.963 | 0.021 |

The PCs are interpreted as follows:

- *PC1:* CBGR, LOC, FanIn and FanOut are coupling and size metrics. We have size and coupling metrics in this dimension. This shows that there are classes with high internal methods i.e. methods defined in the class and external methods i.e. methods called by the class. This means coupling is related to number of methods and attributes in the class.
- *PC2:* IFC measure the total level of information flow of a module and reflect the degree of flow complexity among classes.

### c) *Experiments*

At the first step, we collect the source code of the applications for the study i.e. JEdit 4.2, Velocity 1.4 and Velocity 1.6 application. We enter the source code for each application to a developed C# tool in order to generate call graph model for each application. The developed tool computes the call graph based metrics for each extracted function. Then compute the same metrics to classes and output the results into comma separate value file that represent the data set to be tested. The next step is refining the data set with bug report related to each application under investigation.

Finally, evaluate the value of the metrics in terms of bug and defect detection the format of the data set should be ARRF file as the classifier algorithms such as J48 and M5P algorithm accepts only the files with that format. The accuracy is calculated with tenfold cross validation. The attributes of the file listed in the Figure 2.

@attribute “Number” “numeric”
@attribute “LOC” “numeric”
@attribute “Fan In” “numeric”
@attribute “Fan Out” “numeric”
@attribute “CGBR” “numeric”
@attribute “IFC” “numeric”

*Figure 2 :* Data set attributes

The attribute bug is classified into three categories based on the number of bugs for each class as shown in Table 3.

*Table 3 :* Bugs categories

| Bug Categories | Metric Matrix |
|---|---|
| One | VL = 0 error / L = 1 error / M = 2 error / H = 3 errors / VH => 3 errors |
| Two | L = 0 error / M = 1-2 errors / H => 2 errors |
| Three | False = no error / True = error exist |

The experiments result shows that there is an obvious correlation between the call graphs based metrics, bugs and defects of the application. The result of all the nine experiments is summarizing by Table 4.

*Table 4 :* The experiments results summary in terms of bug categories

| Bug Category Application Name | Category 1 | Category 2 | Category 3 |
|---|---|---|---|
| JEdit 4.2 | 81.34 % | 80.84 % | 86.93 % |
| Velocity 1.4 | 60.67 % | 72.07 % | 80.04 % |
| Velocity 1.6 | 66.59 % | 67.36 % | 73.83 % |

The correlation between bug and the call graph based metrics will be high when we split the bug class into small number of categories, like category three that split the bug class into two categories. So we take category three as criteria to compare the J48 classifier on the applications under investigation output to other classifier output such as logistic model trees (LMT) and support vector machine (SMO) classifier algorithm.

The results of three classifier algorithms have approximately similar values and we conclude that correlation is very high between the call graph metrics and bugs of the application under investigation as shown in Table 5.

*Table 5 :* The experiments result summary in terms of algorithm types

| Classifier Algorithm Application Name | J84 | LMT | SMO |
|---|---|---|---|
| J Edit 4.2 | 86.142 % | 84.926 % | 82.547 % |
| Velocity 1.4 | 80.928 % | 80.364 % | 75.723 % |
| Velocity 1.6 | 72.152 % | 71.029 % | 66.487 % |

Finally, we make some normalization to our data set by excluding the non public functions such as private and protected functions from the computation of the call graph metrics for the applications under investigation and the results of analysis is shown in Table 6.

*Table 6 :* The experiments results summary data set excluding non-public functions

| Classifier Algorithm Application Name | J84 | LMT | SMO |
|---|---|---|---|
| JEdit 4.2 | 86.924 % | 85.196 % | 83.537 % |
| Velocity 1.4 | 85.918 % | 88.364 % | 75.783 % |
| Velocity 1.6 | 72.125 % | 70.709 % | 67.467 % |

The results of three classifier algorithm are approximately have similar values where that leads us to conclude that correlation is very high between the call graph metrics that computed without non public functions and bugs of the application under investigation as shown in Table 6.

After comparing the results of Table 5 and Table 6, we show that excluding the non-public functions such as private and protected functions in order to compute the call graph based metrics for the classes of the application under investigation will raise the percentage of the supposed correlation between call graphs based metrics and bugs.

## V. Conclusion

In this paper, we present the effectiveness and the power of call graph based metrics in prediction and detection the defects in software through our developed tool. We choose three applications J Edit 4.2, Velocity 1.4 and Velocity 1.6. We extract the call graph based metrics such LOC, Fan In, Fan Out, SGBR and IFC from the selected applications and evaluate their correlation according to many categories of bugs for the applications. By all these experiments we discover that how much the extracted call graph metrics are necessary and important in lightening the expensive and time consumer obstacles and problems of software that may arise after delivery phase. Therefore, it will be more effective to predict them and find their solutions earlier before they occur at any time.

The results of our research improve the hypothesis of correlation between call graph based metrics and bugs in software design. The highest percentage of correlation was shown in results of the analysis J Edit 4.2 application using J48 algorithm classifier with metric correlation 86%, while the metric correlation resulted in analysis velocity application with its versions 1.4 and 1.6 was 85% and 72% respectively. In addition, the results show that correlation between bugs and the call graph based metrics will be high when we split the bug class into small number. In addition, the results show that excluding non-public functions such as private and protected functions in order to compute the call graph based metrics for the classes of the application under investigation will raise the percentage of the supposed correlation.

By this approach, we proved that call based metrics are appropriate criteria for helping the maintenance and developing stages to be more effective and less costly at the same time for the systems those are very complex and hardly to understand.

## References Références Referencias

1. J. Dollner and J. Bohnet "Visual exploration of function call graphs for feature location in complex software systems" Proceedings of 2010 ACM symposium on Software visualization vol. 1 (2010) pp. 95 - 104.
2. D. Reniers, A. Telea, O. Ersoy and H. Hoogendorp "Extraction and visualization of call dependencies for large C/C++ code bases: A comparative study" 2011 7th IEEE International Workshop on Visualizing

Software for Understanding and Analysis (2011) pp. 81 - 88.

3. M. Jahromi and E. Honar "A framework for call graph construction" Student thesis At School of Computer Science, Physics and Mathematics (2012).
4. F. B. Abreu and A. L. Baroni "A formal library for aiding metrics extraction" 8th International Workshop on Object Oriented Reengineering (2013) Dramstandt, Germany.
5. A. Bernstein, M. Pinzger and P. Knab "Predicting defect densities in source code files with decision tree learners" Proceedings of the 2011 international workshop on Mining software repositories (2011) pp. 22 - 23, Shanghai, China.
6. T. Ball and N. Nagappan "Use of relative code churn measures to predict system defect density" Proceedings of the 37th international conference on Software engineering (2012) pp. 284 - 292, St. Louis, MO, USA.
7. S. Usmani and N. Azeem "Defect prediction leads to high quality product" Journal of Software Engineering and Applications, vol. 4 (2011) pp. 639 - 645.
8. A. Khare, P. Batra and M. Kaur "Static analysis and run-time coupling metrics" International Journal of Information Technology and Knowledge Management, vol. 6 (2013) pp. 707 - 710.
9. P. Darbyshire and W. Prins "Call graph based program analysis with .Net" In Procs of the IRMA International Conference (2012) pp. 794 - 798.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Intuitionistic Partition based Conceptual Granulation Topic-Term Modeling

By D. Malathi & S. Valarmathy

*Bannari Amman Institute of Technology, India*

*Abstract-* Document Analysis represented in vector space model is often used in information retrieval, topic analysis, and automatic classification. However, it hardly deals with fuzzy information and decision-making problems. To account this, Intuitionistic partition based cosine similarity measure between topic/terms and correlation between document/topic are proposed for evaluation. Conceptual granulation is emphasized in the decision matrix expressed conventionally as tf-idf. A local clustering of topic-terms and document-topics results in comparing dependent terms with membership degree using cosine similarity measure and correlation. A preprocessing of documents with intuitionistic fuzzy sets results in efficient classification of large corpus. But it depends on the datasets chosen. The proposed method effectively works well with large sized categorized corpus.

*Keywords: document analysis, intuitionistic fuzzy, topic modeling.*

*GJCST-C Classification: K.6.3*

INTUITIONISTICPARTITIONBASEDCONCEPTUALGRANULATIONTOPIC-TERMMODELING

*Strictly as per the compliance and regulations of:*

Open Association of Research Society, USA
Research | Diversity | Ethics

# Intuitionistic Partition based Conceptual Granulation Topic-Term Modeling

D. Malathi [α] & S. Valarmathy [σ]

*Abstract-* Document Analysis represented in vector space model is often used in information retrieval, topic analysis, and automatic classification. However, it hardly deals with fuzzy information and decision-making problems. To account this, Intuitionistic partition based cosine similarity measure between topic/terms and correlation between document/topic are proposed for evaluation. Conceptual granulation is emphasized in the decision matrix expressed conventionally as tf-idf. A local clustering of topic-terms and document-topics results in comparing dependent terms with membership degree using cosine similarity measure and correlation. A preprocessing of documents with intuitionistic fuzzy sets results in efficient classification of large corpus. But it depends on the datasets chosen. The proposed method effectively works well with large sized categorized corpus.

*Keywords: document analysis, intuitionistic fuzzy, topic modeling.*

## I. Introduction

Document model in the information retrieval has three main components, namely Text Pre-processor, Topic Extractor and Corpus category-zation. These components are integrated to deploy knowledge extraction in information system. In spite of this, the growing data and its knowledge recognition complications have considerably encouraging the extensions of machine learning algorithms.

### *a) Document Model*

The text document Modeling is observed as latent topics model. Various prominent approaches in machine learning are used to study the model. Document model is a mixture of topics [4]. Topics are inferred by the collection of correlated words. But unsupervised learning perspective is the pulse of bubbling out the topics. By modeling, varieties of mining range can be established with various subjects. The models try to observe the likely documents and tend to focus on topics. But document models are discriminant because of random words due to linguistic factors such as synonym, hyponym, Polysemy, etc.

### *b) Text Pre-processor*

The functionalities essential for machine learning of document are document pre-processing and corpus representation. Stop words removal, word stemming, filtering to exclude certain words, are done within each document. This process is called pre-processing of documents. Obtained vocabulary is put up in the word-document matrix which is generally called as bag-of-words model. The document representations may be in binary (0, for nonoccurrence and 1 for occurrence of each term in a document), term frequency (tij - number of occurrence of ith word in jth document) and term frequency inverse document frequency (probable occurrence of tij' – distribution of ith word in jth document). Obtained data in this stage is huge in dimension, and lot of techniques [15] have been proposed for dimension reduction.

### *c) Topic Extractor*

A topic model is a probabilistic model that can be considered as a mixture of topics, represented by probability distributions of words in a document. The latent variables or topics are the inferring components of this model. The main objective is to learn from documents the distribution of the underlying topics in a given corpus. Topic model is Text corpora representation by a co-occurrence matrix of words and documents. The probabilistic latent semantic analysis (PLSA) model [10] uses probability of words with given topics and probability of topics in a document, to build a topic model. The Latent Dirichlet Allocation (LDA) model [1], is another probabilistic approach which ties the parameters of all documents through hierarchical generative model.

### *d) Corpus Categorization*

Text Categorization is a classical application of Text Mining [19], and is used in email filters, social tagging and automatic labeling of documents in business libraries. Text mining applications in research and business intelligence include, latent semantic analysis techniques in bioinformatics automatic investigation of jurisdictions plagiarism detection in universities and publishing houses, cross-language information retrieval, spam filters learning, help desk inquiries, measuring customer preferences by analyzing qualitative interviews, automatic grading, fraud detection or parsing social network for ideas of new products [9].

## II. Literature Support

The theory of fuzzy set is Consider as a degree of membership assigned to each element, where the degree of non-membership is just automatically equal to

*Author α σ: Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India. e-mails: malathisubbu@gmail.com, artmathy@gmail.com*

its complement. However, human interpretation often does not express the corresponding degree of non-membership as the complement to 1. So, Atanassov [1][2][3] introduced the concept of intuitionistic fuzzy set that is meant to reflect the fact that the degree of non-membership is not always equal to 1 minus degree of membership, but there may be some hesitation degree.

Intuitionistic fuzzy set is a generalized constructive logic applied in fuzzy set. It is defined on a $X$ of objects, with each object $x$ is described by the degrees of membership and non-membership to a certain property,

$$\{(x, \mu_A(x), \nu_A(x)), x \in X\} \qquad (1)$$

$\mu_A(x)$ represents the degree x belongs to the set A and $\nu_A(x)$ represents the degree x does not belongs to the set $A$. The model is defined by the restriction

$$0 \le \mu_A(x) + \nu_A(x) \le 1 \qquad \forall x \in X \qquad (2)$$

Therefore the degree of non determinacy of the object $x$ with respect to the intuitionistic fuzzy set $A$ is imposed as,

$$\pi_A(x) = \mu_A(x) + \nu_A(x) \quad \forall x \in X \qquad (3)$$

The model is well suited to represent a classification problem with high dimension. The confusion matrix of high dimension can be probably reduced to concept matrix of low dimension. The similarity measures [14] and distance measures [21][20] between two intuitionistic fuzzy sets can be applied in pattern recognition.

In this paper, a Partition based approach [16] inspired by Hierarchical segmentation [8] and topic based segmentation [6] are extended using Intuitionistic fuzzy set approach [23] for local centralization of conceptual words. The intuitionistic fuzzy set theory is applied in conceptual term/topic detection. A cosine similarity and correlation are taken into for defining membership degree and the non-membership degree respectively. The results using this measure found better with respect to the dataset chosen. In literature a intuitionistic fuzzy representation of images for clustering [18] [12] by utilizing a novel similarity metric are defined. But a minimal support is extended for text classification. So, a local centralization of conceptual terms using Intuitionistic logical clustering has been applied in the work.

## III. Proposed Model - Intuitionistic Partition based Concept Granulation (IPCG)

Intuitionistic logic is a natural deduction system [13],that have introduction rules $\mu$ and elimination rules $\nu$ for the logical connectives and quantifiers. The document classification system needs conceptual terms $(\mu)$, non deterministic terms or noises $(\nu)$ with logics and reasons to quantify concept granules.

Let A be a tf-idf matrix of $nXm$ represents corpus. Each value is associated to

- Set of terms representing the membership of domain $\mu_A(x)$
- Term representing the non membership of domain $\nu_A(x)$

```
Algorithm: IPCG
For each document {
        Lowercase, numbers, special characters from document
        Remove stop list words from document
      Split document into k partitions
      For each segment {
      Find frequency of words
      Prepare matrix with each segment as row and words as
columns
      Include non zero frequency as member
      Cosine similarity distance between each segments
calculated
      Discard the segment with least distance }
   Single row or vector of a document has been found
   Intuitionistic Correlation to include conceptual terms in topic
   Classify the document and find entropy }
```

The intuitionistic fuzzy set A is generated by

$$A = \{w_{ij}, \mu_i(w_{ij}), \nu_i(w_{ij})\} \quad \text{where} \quad 0 < w_{ij} < 1 \qquad (4)$$

The similarity between words and on a topic is calculated by the cosine measure. Each document vector is normalized with the weight and length of terms in $k$ partitions. Then the optimal term $w_{ij}$ [16] should be picked from the non sparse term of $k$ partitions. i.e. $\mu_A \in w_{ik}$.

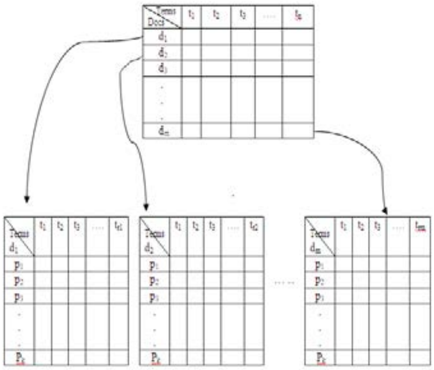


*Figure 1 :* Partition Model

{ri<=n (i.e. r is random or varies from document to document)(where i=1,2,…m),
k = no. of partitions or segments}

$$\mu_i = \frac{1}{|w_{ik}|}\sum w_{ik} \quad \text{where } w_{ik} > 0 \text{ and } |w_{ik}| >= k/2 \qquad (5)$$

The intuitionistic angular or cosine similarity [22] measure between the m terms in a partitioned set is given as follows:

$$C(A,B) = \frac{\sum_{i=1}^{m} \mu_A(x_i)\mu_B(x_i)}{\sqrt{\sum_{i=1}^{m} \mu_A^2(x_i)}\sqrt{\sum_{i=1}^{m} \mu_B^2(x_i)}} \qquad (6)$$

The intuitionistic correlation [7] of rows all fuzzy numbers are included from the samples of tf-idf (Partition Model). The crisp set is modified intuitionistically with the sample mean and variance of membership function as:

$$CR_I(A,B) = \frac{\left(\sum_{i=1}^{n} (\mu_A(x_i) - \overline{\mu}_A)(\mu_B(x_i) - \overline{\mu}_B)\right)}{\sqrt{\sum_{i=1}^{n} (\mu_A(x_i) - \overline{\mu}_A)}\sqrt{\sum_{i=1}^{n} (\mu_B(x_i) - \overline{\mu}_B)}} \qquad (7)$$

The effectiveness of the intuitionistic classification of corpus is approximately studied and analyzed using the following entropy [22] specifically used for Intuitionist Fuzzy Set 'A'.

$$E = \frac{1}{n}\sum_{i=1}^{n} \frac{\min(\mu_A(x_i), v_A(x_i)) + \pi_A(x_i)}{\max(\mu_A(x_i), v_A(x_i)) + \pi_A(x_i)} \qquad (8)$$

## IV. Datasets

*a) Newspaper Article collection*

The newspaper articles under different topics are collected. The categories are marked. The training and testing documents are randomly chosen. The growing social media made essential to include newspaper article collection to include in this work. News are generally categorized by topic area ("politics," "business," etc.) written in clear, correct, "objective," and somewhat schematized language [5]. This would pave way to extend the research towards social networking and marketing. The collection includes about 780 documents with 25 categories. All new social relevant topics ("mobile","opinion", etc.) are included for categorizing.

*b) Reuters-21578 Data Set*

The Reuters-21578 Data Set collection provides a classification task with challenging properties. There are multiple categories, the categories are overlapping and non exhaustive, and there are relationships among the categories. There are interesting possibilities for the use of domain knowledge. There are many possible feature sets that can be extracted from the text, and most plausible feature/example matrices are large and sparse [11].

*c) Movie Review Dataset*

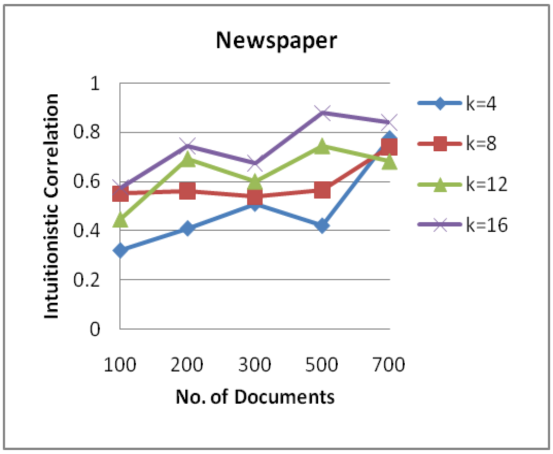
The Movie Review Dataset, Polarity dataset v0.9 with 900 positive and 900 negative reviews is used. Using movie reviews as data, the problem of classifying documents using standard machine learning techniques definitively outperform human-produced baselines processed reviews [17]. The training cases are chosen randomly from each class about 100 documents. Which means about 500 cases are considered for training.

## V. Results and Analysis

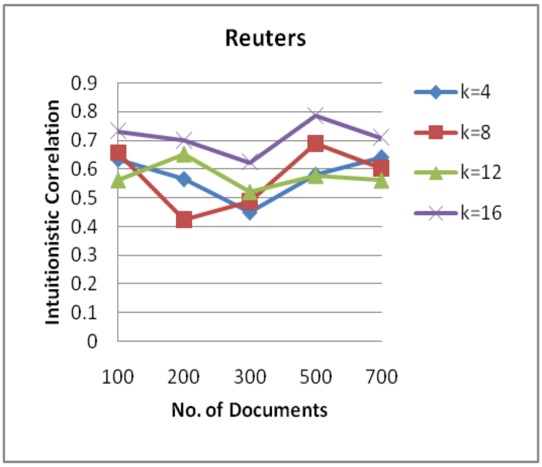
The machine learning classification methods, such as Bayesian, Naïve Bayes, J48, Support Vector Machines, LMT are strong enough to support classifications.

In the case of concept granulation in document classification, the feature selection is fine tuned to achieve categories strictly connected to the human perception. Before imposing the features into the classifier, some form of selection must be chosen. The proposed method, selects the features according to the intuitionist logic. The features tf-idf matrix has been transformed into intuition based feature model. The proposed approach is modeled as a probability distribution over the set of Topic/Words represented by the vocabulary. These distributions are sampled from multi-nominal distributions.

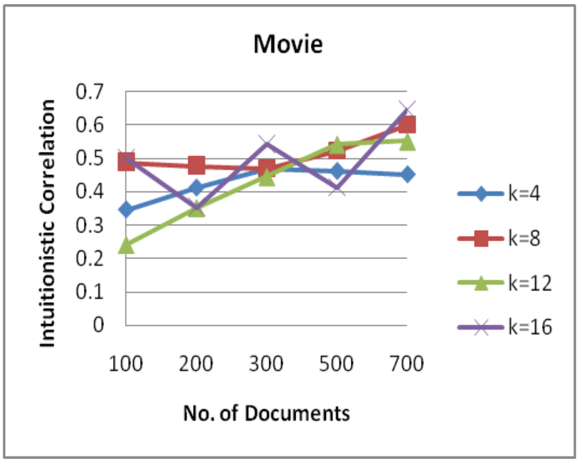
The proposed Concept Granulation Using Intuitionistic Partition Based Classification Model is implemented administered in the Java based system and analyzed for its significance. The intuitionistic correlation is applied to the specified datasets. In which the chosen dataset and the partitions play the very important role in finding the result of the model. The tf-idf-IP is favorable for Reuter dataset than for Newspaper and Movies. This is represented in the Figures 2(a) (b) (c). Reuters in which documents are well organized behaves highly significant to the model. In Newspaper collection, the documents are synthetically collected and organized. But due to the nature of news along with the temporal parameters, it is moderately supported by the model. The least support is favored by the movie dataset. This is due to the heterogeneity of the documents/terms/topics.



(a)



(b)



(c)

*Figure 2 :* Intuitionistic correlation Vs The number of training documents

The perplexity is depicted in Figure 3 and Table1. So the analysis can be interpreted or inferred in the following ways:

Intuitionistic approach is in favor of the classified documents or corpus chosen

Partition plays the important role in the proposed model. Out of four types of partition, k=8 plays a smoothened strong support for the proposed model

k=16, the highest partition yield only a very moderate result and more confusions.

k=4, the least partition model yield the smooth but less significant support for all the datasets.

k=8, yield the partially smooth but supportive significant for the movie dataset. (Than other partitions)

The results are focused to average training datasets and micro f-measure (Table 2) to show up the IPCG performs better with dimension reduction for categorization of corpus. Every datasets chosen for analysis behaves to the pull and push of various stages of the proposed model.

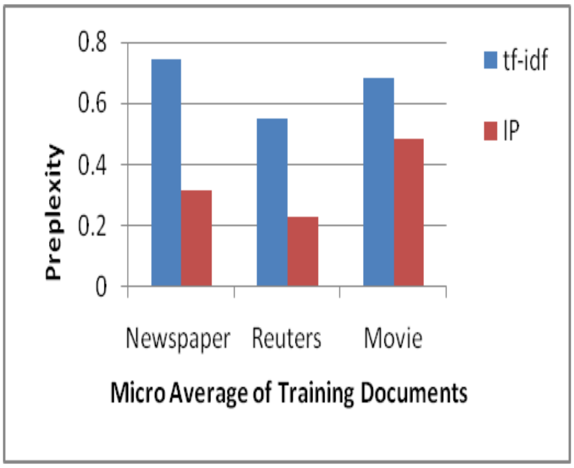


*Figure 3 :* Confusions in classification

*Table 1 :* Confusions in classification

| Training with 300 Doc | Dimension Reduction | Perplexity | Correlation |
|---|---|---|---|
| Newspaper | 26% | 0.231 | 0.582 |
| Reuters | 22% | 0.311 | 0.520 |
| Movie | 16% | 0.483 | 0.480 |

*Table 2 :* Micro Evaluation of F-measure with average training sets

| Dataset | tf-idf | | | IPCG | | |
|---|---|---|---|---|---|---|
| Classifiers | Reuters | News Paper | Movie | Reuters | News Paper | Movie |
| **SVM** | 0.482 | 0.422 | 0.321 | 0.844 | 0.841 | 0.799 |
| **NB** | 0.401 | 0.369 | 0.297 | 0.872 | 0.834 | 0.810 |
| **J48** | 0.400 | 0.399 | 0.381 | 0.798 | 0.797 | 0.784 |
| **Bayes'** | 0.541 | 0.411 | 0.399 | 0.831 | 0.854 | 0.829 |
| **LMT** | 0.442 | 0.541 | 0.587 | 0.878 | 0.798 | 0.722 |

## VI. Conclusions

In this paper, we have proposed a intuitionistic partition based concept granulation topic-term model for a nominal tf-idf vector space model which is often used in information retrieval, topic analysis, and automatic classification. The cosine distance and correlation treatment to the tf-idf reduces the dimension and improves the efficiency of bag of words/terms in topics. However, it is priory treated using the intuitionistic partition for fitting the model into decision-making problems. To account this, Intuitionistic partition based cosine similarity measure between topic/terms and correlation between document/topic are included. The proposed fuzzy model is tailored with normal combinational approach to fetch intuitionistic fuzzy crisp set. Yet, it is observed the model is well behaving and promising for the categorized documents and not so bad support for the low inference corpus collections like movie review. So, this make us clear that the social media documents should be specially treated before introducing this model. It is felt that aggregation of social media topic-terms is needed. This is taken for future work or extension of the proposed work.

## References Références Referencias

1. D. M. Blei, A. Y. Ng and M. I. Jordan. Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3(2003), 993-1022.
2. T. Hofmann. Probabilistic Latent Semantic Analysis. *In Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99),* San Francisco, CA, Morgan Kaufmann, (1999), 289–329.
3. F. Sebastiani, Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, 34(2002), 1-47. doi:10.1145/505282.505283.
4. I. Feinerer, K. Hornik and D. Meyer, *Journal of Statistical Software*, 25, (2008), 1-54.
5. K.T. Atanassov, Intuitionistic Fuzzy Sets, Theory, and Applications, Series in Fuzziness and Soft Computing, *Phisica-Verlag*, 1999.
6. K.T. Atanassov, Intuitionistic Fuzzy Set*, Fuzzy Sets System*, (1986) 87–97.
7. K.T. Atanassov, S. Stoeva, Intuitionistic Fuzzy Set, *Polish Symposium on Interval and Fuzzy Mathematics, Poznan*, (1993), 23–26.
8. D. Li and C. Cheng, New Similarity Measures Of Intuitionistic Fuzzy Sets And Application To Pattern Recognition, *Pattern Recognition Letter*, 23(2002) 221–225.
9. E. Szmidt and J. Kacprzyk, Entropy for Intuitionistic Fuzzy Set, *Fuzzy Sets System*, 118 (2001), 467–477.
10. E. Szmidt and J. Kacprzy k, Distance Between Intuitionistic Fuzzy Set, *Fuzzy Sets System*, 114(2000), 505–518.
11. D. Malathi, S. Valarmathy, Domain Classifier using Conceptual Granulation and Equal Partition Approach, *Indian Journal of Engineering, Science and Technology*, 7(2013), 39-43.
12. J. T. Chien and C. H. Chueh, Topic-Based Hierarchical Segmentation, *IEEE Transactions on Audio, Speech and Language Processing*, 20(2012), 55-66.
13. T. Brants, F. Chen and I. Tsochantaridis, Topic-based document segmentation with Probabilistic Latent Semantic Analysis, *in the proceeding of International Conference on Information and Knowledge Management*, (2002), 211–218.

14. Z. Xu, J. Chen and J. Wu. Clustering Algorithm for Intuitionistic Fuzzy Sets. Information Sciences, 178(2008), 3775-3790.
15. N. Pelekis, D. K. Iakovidis, E. K. Evangelos and I. Kopanakis. Fuzzy Clustering of Intuitionistic Fuzzy Data, *International Journal of Business Intelligence and Data Mining*, 3(1), pp. 45-65.
16. D. K. Iakovidis, N. Pelekis, E. K. Evangelos and I. Kopanakis, Intuitionistic Fuzzy Clustering with Applications in Computer Vision. *Advanced Concepts for Intelligent Vision Systems, Lecture Notes in Computer Science,* 5259(2008), 764-774.
17. V. H. Le, C. H. Nguyen and F. Liu, Semantics and Aggregation of Linguistic Information, Based on Hedge Algebras, *The 3rd International Conference on Knowledge, Information, and Creativity Support Systems*, (2013).
18. J. Ye, Multicriteria Decision-making Method Based on a Cosine Similarity Measure between Trapezoidal Fuzzy Numbers, *International Journal of Engineering, Science and Technology,* 3(2011), 272-278.
19. D. A. Chiang, and N. P. Lin, Correlation of fuzzy sets, *Fuzzy Sets and Systems*, 102(1999), 221-226.
20. B. Berendt, Text Mining for News and Blogs Analysis. *In C. Sammut, & G. I. Webb, Encyclopedia of Machine learning*, London: Springer. (2010), 968-972.
21. http://www.daviddlewis.com/resources/testcollections/reuters21578.
22. B. Pang, L. Lee and S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, *Proceedings of EMNLP*, (2002), 79-86.
23. D. Malathi and S. Valarmathy, A Comprehensive Survey on Dimension Reduction Techniques for Concept Extraction from a Large Corpus, *International Journal of Computing Information Systems,* 3(2011), 1-6.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# An Advanced Clustering Algorithm (ACA) for Clustering Large Data Set to Achieve High Dimensionality

By Aman Toor

*Abstract-* Cluster analysis method is one of the main analytical methods in data mining; this method of clustering algorithm will influence the clustering results directly. This paper proposes an Advanced Clustering Algorithm in order to solve this question, requiring a simple data structure to store some information [1] in every iteration, which is to be used in the next iteration. The Advanced Clustering Algorithm method avoids computing the distance of each data object to the cluster centers repeat, saving the running time. Experimental results show that the Advanced Clustering Algorithm method can effectively improve the speed of clustering and accuracy, reducing the computational complexity of the traditional algorithm. This paper includes Advanced Clustering Algorithm (ACA) and describes the experimental results and conclusions through experimenting with academic data sets.

*Keywords: ACA, SOM, K-MEANS, HAC, clustering, large data set, high dimensionality, cluster analysis.*

*GJCST-C Classification: H.3.3*

ANADVANCEDCLUSTERINGALGORITHMACAFORCLUSTERINGLARGEDATASETTOACHIEVEHIGHDIMENSIONALITY

*Strictly as per the compliance and regulations of:*

# An Advanced Clustering Algorithm (ACA) for Clustering Large Data Set to Achieve High Dimensionality

Aman Toor

*Abstract-* Cluster analysis method is one of the main analytical methods in data mining; this method of clustering algorithm will influence the clustering results directly. This paper proposes an Advanced Clustering Algorithm in order to solve this question, requiring a simple data structure to store some information [1] in every iteration, which is to be used in the next iteration. The Advanced Clustering Algorithm method avoids computing the distance of each data object to the cluster centers repeat, saving the running time. Experimental results show that the Advanced Clustering Algorithm method can effectively improve the speed of clustering and accuracy, reducing the computational complexity of the traditional algorithm. This paper includes Advanced Clustering Algorithm (ACA) and describes the experimental results and conclusions through experimenting with academic data sets.

*Keywords: ACA, SOM, K-MEANS, HAC, clustering, large data set, high dimensionality, cluster analysis.*

## I. Introduction

Clustering is the process of organizing data objects into a set of disjoint classes called Clusters. Clustering is an Unsupervised Clustering technique of Classification. Classification refers to a technique that assigns data objects to a set of classes. Unsupervised means that clustering does not depends upon predefined classes while clustering the data objects. Formally, given a set of dimensional points and a function that gives the distance between two points , we are required to compute cluster centers, such that the points falling in the same cluster are similar and points that are in different cluster are dissimilar. Most of the initial clustering techniques were developed by statistics or pattern recognition communities, where the goal was to cluster a modest number of data instances. However, within the data mining community, the focus has been on clustering large datasets [2]. Developing clustering algorithms to effectively and efficiently cluster rapidly growing datasets has been identified as an important challenge.

A number of clustering algorithms have been proposed to solve clustering problems. One of the most popular clustering methods are K-Means, SOM, HCA. Their shortcomings are discussed below.

The standard k-means algorithm needs to calculate the distance from the each data object to all the centers of k clusters when it executes the iteration each time, which takes up a lot of execution time especially for large-capacity databases. In K-Means algorithm initial cluster centers are produced arbitrary, it does not promise to produce the peculiar clustering results. Efficiency of original k-means algorithm is heavily rely on the initial centroid. Initial centroid also has an influence on the number of iterations required while running the original K-Means algorithm. Computational Complexity of K-Means algorithm is very high and does not provide high quality clusters when it comes to cluster High dimensional data set.

Kohonon's SOMs are a type of unsupervised learning. The goal is to discover some underlying structure of the data. SOM algorithm is computationally expensive. Large quantity of good quality representative training data required. No generally accepted measure of 'quality' of a SOM e.g. Average quantization error (how well the data is classified). Every SOM is different therefore we must be careful what conclusions we draw from our results. SOM is non-deterministic and can and will produce different results in different run.

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC. [6] Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached. This algorithm is sensitive to outliers and sometimes it is difficult to identify the correct number of clusters from Dendrogram. [7]

Various methods have been proposed in literature but it have been analyzed that the K-Means, SOM, HCA fails to give optimum result when it comes to clustering high dimensional data set because their complexity tends to make things more difficult when number of dimensions are added. In data mining this problem is known as *"Curse of Dimensionality*". This research will deal the problem of high dimensionality and large data set.

A large number of algorithms had been proposed till date, each of them address some specific

*Author: e-mail: er.amantoor@gmail.com*

requirement. There does not exist a single algorithm which can adequately handle all sorts of requirement. This makes a great challenge for the user to do selection among the available algorithm for specific task. To cope with this problem, a new algorithm is going to be proposed in this research that is named as "*Advanced Clustering Algorithm".*

This paper is organized s follows. Section 2 presents an overview of ACA. Section 3 introduces proposed method. Section 4 describes about the time complexity of proposed method. Section 5 experimentally demonstrates the performance of proposed method. And the final Section 6 describes the conclusion.

## II. Advanced Clustering Algorithm

Experimental results have shown Kohonon's SOM is superlative clustering algorithm among K-means, HCA [8]. For the shortcomings of the above SOM algorithm, this paper presents an Advanced Clustering Algorithm method. The main idea of algorithm is to set two simple data structures to retain the labels of cluster and the distance of all the data objects to the nearest cluster during the each iteration that can be used in next iteration. We calculate the distance between the current data object and the new cluster center, if the computed distance is smaller than or equal to the distance to the old center, the data object stays in it's cluster that was assigned to in previous iteration. Therefore, there is no need to calculate the distance from this data object to the other k- 1 clustering centers, saving the calculative time to the k-1 cluster centers. Otherwise, we must calculate the distance from the current data object to all k cluster centers, and find the nearest cluster center and assign this point to the nearest cluster center. And then we separately record the label of nearest cluster center and the distance to it's center. Because in each iteration some data points still remain in the original cluster, it means that some parts of the data points will not be calculated, saving a total time of calculating the distance, thereby enhancing the efficiency of the algorithm.

## III. Proposed Algorithm

The process of the Advanced Clustering algorithm is described as follows: Input: The number of desired clusters k, and a database D= {$d_1$, $d_2$, $d_n$} containing n data objects. Output: A set of k clusters.

1. Draw multiple sub-samples {SI, S2, . . . ,Sj } from the original dataset.
2. Repeat step 3 for m=I to i
3. Apply combined approach for sub sample.
4. Compute centroid
5. Choose minimum of minimum distance from cluster center criteria
6. Now apply new calculation again on dataset S for k1 clusters.
7. Combine two nearest clusters into one cluster and recalculate the new cluster center for the combined cluster until the number of clusters reduces into *k.*

## IV. Time Complexity

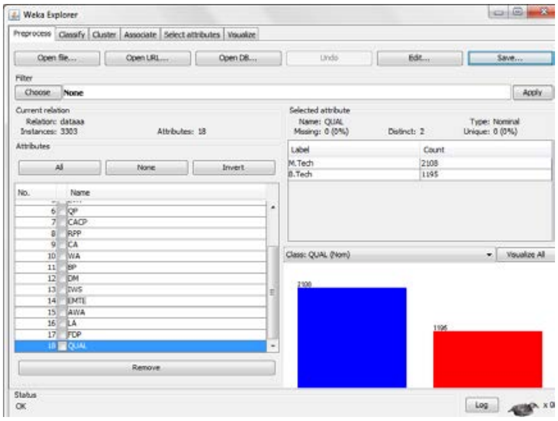
This paper proposes an Advanced Clustering Algorithm, to obtain the initial cluster, time complexity of the advanced algorithm is O (nk). Here some data points remain in the original clusters, while the others move to other clusters. If the data point retains in the original cluster, this needs O (1), else O (k). With the convergence of clustering algorithm, the number of data points moved from their cluster will reduce. If half of the data points move from their cluster, the time complexity is O (nk/2). Hence the total time complexity is O (nk). While the standard k-means clustering algorithm require O(nkt). So the proposed algorithm in this paper can effectively improve the speed of clustering and reduce the computational complexity. But the Advanced k-means algorithm requires the pre estimated the number of clusters, k, which is the same to the standard k-means algorithm. If you want to get to the optimal solution, you must test the different value of k.

## V. Experimental Results

This paper selects academic data set repository of machine learning databases to test the efficiency of the advanced algorithm and the standard algorithms. Two simulated experiments have been carried out to demonstrate the performance of the Advanced in this paper. This algorithm has also been applied to the clustering of real datasets. In two experiments, time taken for each experiment is computed. The same data set is given as input to the standard algorithm and the Advanced Clustering Algorithm. Experiments compare Advanced Clustering Algorithm with the standard algorithm in terms of the total execution time of clusters and their accuracy. Experimental operating system is Window 8, program language is java. This paper uses academic activities as the test datasets and gives a brief description of the datasets used in experiment evaluation. Table 1 shows some characteristics of the datasets.

*Table 1 :* Data Set Size

| Dataset | Number of attributes | Number of records |
|---|---|---|
| Academic Activities | 15 | 5504 |


Weka Explorer

Filter: None

Current relation
Relation: dataaa
Instances: 3303
Attributes: 18

Selected attribute
Name: QUAL
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

| Label | Count |
|---|---|
| M.Tech | 2108 |
| B.Tech | 1195 |

| No. | Name |
|---|---|
| 6 | QP |
| 7 | CACP |
| 8 | RPP |
| 9 | CA |
| 10 | WA |
| 11 | BP |
| 12 | DM |
| 13 | IWS |
| 14 | EMTE |
| 15 | AWA |
| 16 | LA |
| 17 | FDP |
| 18 | QUAL |

Class: QUAL (Nom)

Status
OK

*Figure 1 :* Display data set according to class attributes

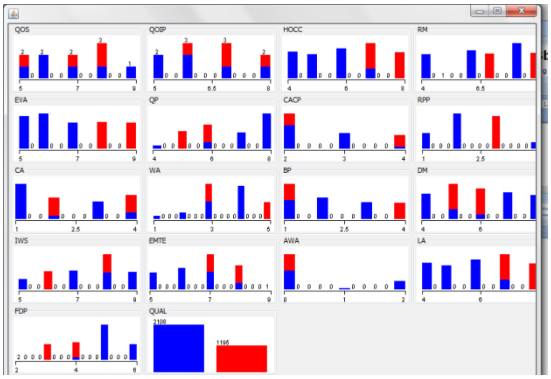
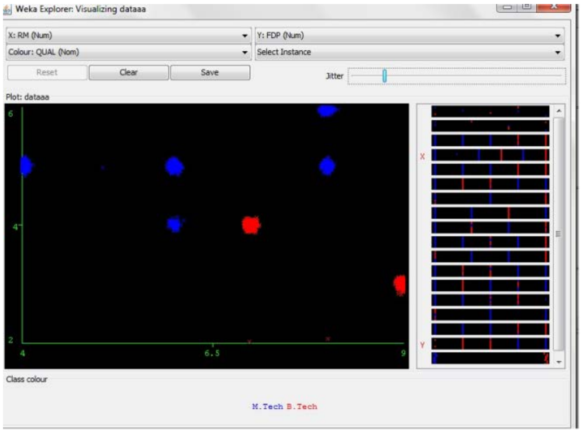


*Figure 2 :* Display All Attributes


Weka Explorer: Visualizing dataaa

X: RM (Num)
Y: FDP (Num)
Colour: QUAL (Nom)
Select Instance
Plot: dataaa

*Figure 3 :* Visualization of scatter plot

*Table 2 :* Analysis between traditional and Advanced Clustering Algorithm

| Parameter | SOM | K-Means | HAC | ECA |
|---|---|---|---|---|
| Error Rate | 0.8189 | 0.8456 | 0.8379 | 0.3672 |
| Execution Time | 297 ms | 1281 ms | 1341 ms | 1000 ms |
| Accessing Time | Fast | Slow | Slow | Very fast |
| Number of Clusters | 6 | 6 | 6 | 4 |

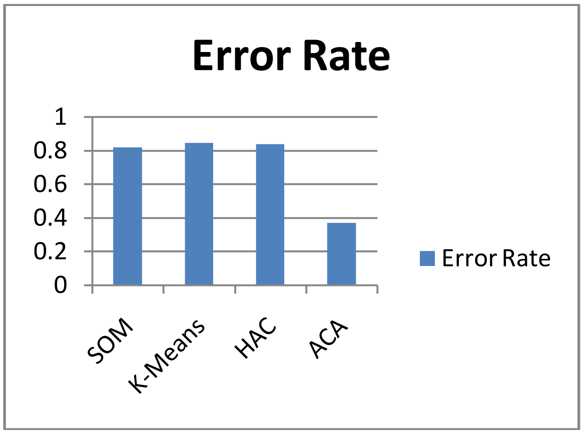


*Figure 4 :* Performance Comparison Based on Error Rate (Quality)

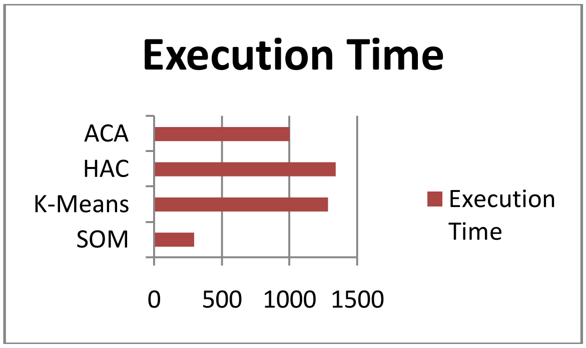


*Figure 5 :* Performance Comparison Based on Execution Time

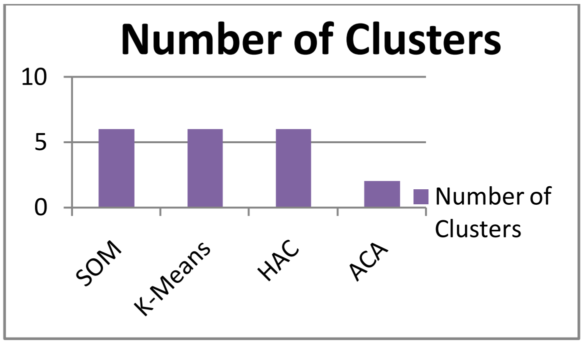


*Figure 6 :* Performance Comparison Based on Number of Clusters

## VI. Conclusion

SOM algorithm is a typical clustering algorithm and it is widely used for clustering large sets of data. This paper elaborates Advanced Clustering Algorithm and analyses the shortcomings of the standard k-means, SOM and HAC clustering algorithm. Because the computational complexity of the standard algorithm is objectionably high owing to the need to reassign the data points a number of times during every iteration, which makes the efficiency of standard clustering is not high. This paper presents a simple and efficient way for assigning data points to clusters. The proposed method in this paper ensures the entire process of clustering in O (nk) time without sacrificing the accuracy of clusters. Experimental results show the Advanced Clustering Algorithm can improve the execution time, quality of SOM algorithm and works well on High Dimensional data set. So the proposed method is feasible.

## References Références Referencias

1. Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29, August 2004.
2. Sun Jigui, Liu Jie, Zhao Lianyu, "Clustering algorithms Research", Journal of Software, Vol 19,No 1, pp.48-61,January 2008.
3. Sun Shibao, Qin Keyun," Research on Modified k-means Data Cluster Algorithm" I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," Computer Engineering, vol.33, No.13, pp.200– 201,July 2007.
4. Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-learningdatabases
5. Fahim A M, Salem A M, Torkey F A, "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University Science A, Vol.10, pp: 1626-1633,July 2006.
6. Zhao YC, Song J. GDILC: A grid-based density isoline clustering algorithm. In: Zhong YX, Cui S, Yang Y, eds. Proc. of theInternet Conf. on Info-Net. Beijing: IEEE Press,2001.140−145.http://ieeexplore.ieee.org/iel5/7719/21161/00982709.pdf
7. Amanpreet Kaur Toor, Amarpreet Singh, " A Survey paper on recent clustering approaches in data mining", Vol 3, Issue 11, November 2013.
8. Amanpreet Kaur Toor, Amarpreet Singh, " Analysis of Clustering Algorithm based on Number of Clusters, error rate, Computation Time and Map Topology on large Data Set", Volume 2, Issue 6, November- December 2013.
9. Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol.2, pp: 283–304, 1998.
10. K.A. Abdul Nazeer, M.P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceeding of the World Congress on Engineering, vol 1, london, July 2009.
11. Fred ALN, Leitão JMN. Partitionalvs hierarchical clustering using a minimum grammar compexity approach. In: Proc. of the SSPR & SPR 2000. LNCS 1876, 2000.193−202.Ohttp://www.sigmod.or-g/dblp/db/conf/sspr/sspr2000.htm
12. Gelbard R, Spiegler I. Hempel's raven paradox: A positive approach to cluster analysis. Computers and Operations Research, 2000, 27(4):305−320.
13. Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proc. of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tucson, 1997.146−151.
14. Ding C, He X. K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization. In: Proc. of the ACM Symp. on Applied Computing. Nicosia: ACM Press, 2004. 584−589.http://www.acm.org/conferences/sac/sac2004/
15. Hinneburg A, Keim D. An efficient approach to clustering in large multimedia databases with noise. In: Agrawal R, Stolorz PE, Piatetsky- Shapiro G, eds. Proc. of the 4th Int'l Conf. on Knowledge Discovery and Data Mining(KDD'98). New York: AAAI Press, 1998.58~65.
16. Zhang T, Ram akrishnan R, .BIRCH: An efficient data clustering method for very large databases. In: Jagadish HV, Mumick IS, eds. Proc. of the 1996 ACM SIGMOD Int'l Conf. on Management of Data. Montreal: ACM Press, 1996. 103~114. [15] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial- temporal data. Data & Knowledge Engineering, 2007, 60(1): 208-221.

Global Journals Inc. (US) Guidelines Handbook 2014

www.GlobalJournals.org

## FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

Global Journals Incorporate (USA) is accredited by Open Association of Research Society (OARS), U.S.A and in turn, awards “FARSC” title to individuals. The 'FARSC' title is accorded to a selected professional after the approval of the Editor-in-Chief/Editorial Board Members/Dean.

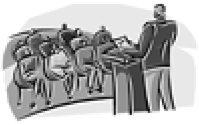
- The “FARSC” is a dignified title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

FARSC accrediting is an honor. It authenticates your research activities. After recognition as FARSC, you can add 'FARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, and Visiting Card etc.

*The following benefits can be availed by you only for next three years from the date of certification:*

FARSC designated members are entitled to avail a 40% discount while publishing their research papers (of a single author) with Global Journals Incorporation (USA), if the same is accepted by Editorial Board/Peer Reviewers. If you are a main author or co-author in case of multiple authors, you will be entitled to avail discount of 10%.

Once FARSC title is accorded, the Fellow is authorized to organize a symposium/seminar/conference on behalf of Global Journal Incorporation (USA).The Fellow can also participate in conference/seminar/symposium organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent.

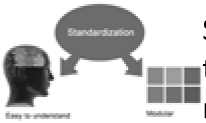


You may join as member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. In addition, it is also desirable that you should organize seminar/symposium/conference at least once.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The FARSC can go through standards of OARS. You can also play vital role if you have any suggestions so that proper amendment can take place to improve the same for the benefit of entire research community.

As FARSC, you will be given a renowned, secure and free professional email address with 100 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.



The FARSC will be eligible for a free application of standardization of their researches. Standardization of research will be subject to acceptability within stipulated norms as the next step after publishing in a journal. We shall depute a team of specialized research professionals who will render their services for elevating your researches to next higher level, which is worldwide open standardization.

The FARSC member can apply for grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A. Once you are designated as FARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria. After certification of all your credentials by OARS, they will be published on your Fellow Profile link on website https://associationofresearch.org which will be helpful to upgrade the dignity.

The FARSC members can avail the benefits of free research podcasting in Global Research Radio with their research documents. After publishing the work, (including published elsewhere worldwide with proper authorization) you can upload your research paper with your recorded voice or you can utilize chargeable services of our professional RJs to record your paper in their voice on request.

The FARSC member also entitled to get the benefits of free research podcasting of their research documents through video clips. We can also streamline your conference videos and display your slides/ online slides and online research video clips at reasonable charges, on request.

The FARSC is eligible to earn from sales proceeds of his/her researches/reference/review Books or literature, while publishing with Global Journals. The FARSC can decide whether he/she would like to publish his/her research in a closed manner. In this case, whenever readers purchase that individual research paper for reading, maximum 60% of its profit earned as royalty by Global Journals, will be credited to his/her bank account. The entire entitled amount will be credited to his/her bank account exceeding limit of minimum fixed balance. There is no minimum time limit for collection. The FARSC member can decide its price and we can help in making the right decision.

The FARSC member is eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get remuneration of 15% of author fees, taken from the author of a respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account.

## MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

The ' MARSC ' title is accorded to a selected professional after the approval of the Editor-in-Chief / Editorial Board Members/Dean.
The “MARSC” is a dignified ornament which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., MARSC or William Walldroff, M.S., MARSC.

MARSC accrediting is an honor. It authenticates your research activities. After becoming MARSC, you can add 'MARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, Visiting Card and Name Plate etc.

*The following benefitscan be availed by you only for next three years from the date of certification.*

MARSC designated members are entitled to avail a 25% discount while publishing their research papers (of a single author) in Global Journals Inc., if the same is accepted by our Editorial Board and Peer Reviewers. If you are a main author or co-author of a group of authors, you will get discount of 10%.

As MARSC, you will be given a renowned, secure and free professional email address with 30 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The MARSC member can apply for approval, grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.

Once you are designated as MARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria.

It is mandatory to read all terms and conditions carefully.

# Auxiliary Memberships

## Institutional Fellow of Open Association of Research Society (USA)-OARS (USA)

Global Journals Incorporation (USA) is accredited by Open Association of Research Society, U.S.A (OARS) and in turn, affiliates research institutions as "Institutional Fellow of Open Association of Research Society" (IFOARS).

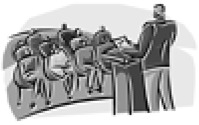
The "FARSC" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

The IFOARS institution is entitled to form a Board comprised of one Chairperson and three to five board members preferably from different streams. The Board will be recognized as "Institutional Board of Open Association of Research Society"-(IBOARS).

*The Institute will be entitled to following benefits:*

The IBOARS can initially review research papers of their institute and recommend them to publish with respective journal of Global Journals. It can also review the papers of other institutions after obtaining our consent. The second review will be done by peer reviewer of Global Journals Incorporation (USA) The Board is at liberty to appoint a peer reviewer with the approval of chairperson after consulting us.

The author fees of such paper may be waived off up to 40%.



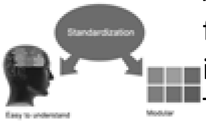
The Global Journals Incorporation (USA) at its discretion can also refer double blind peer reviewed paper at their end to the board for the verification and to get recommendation for final stage of acceptance of publication.

The IBOARS can organize symposium/seminar/conference in their country on behalf of Global Journals Incorporation (USA)-OARS (USA). The terms and conditions can be discussed separately.

The Board can also play vital role by exploring and giving valuable suggestions regarding the Standards of "Open Association of Research Society, U.S.A (OARS)" so that proper amendment can take place for the benefit of entire research community. We shall provide details of particular standard only on receipt of request from the Board.

The board members can also join us as Individual Fellow with 40% discount on total fees applicable to Individual Fellow. They will be entitled to avail all the benefits as declared. Please visit Individual Fellow-sub menu of GlobalJournals.org to have more relevant details.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.



After nomination of your institution as "Institutional Fellow" and constantly functioning successfully for one year, we can consider giving recognition to your institute to function as Regional/Zonal office on our behalf.
The board can also take up the additional allied activities for betterment after our consultation.

**The following entitlements are applicable to individual Fellows:**

Open Association of Research Society, U.S.A (OARS) By-laws states that an individual Fellow may use the designations as applicable, or the corresponding initials. The Credentials of individual Fellow and Associate designations signify that the individual has gained knowledge of the fundamental concepts. One is magnanimous and proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice.

Open Association of Research Society (US)/ Global Journals Incorporation (USA), as described in Corporate Statements, are educational, research publishing and professional membership organizations. Achieving our individual Fellow or Associate status is based mainly on meeting stated educational research requirements.

Disbursement of 40% Royalty earned through Global Journals : Researcher = 50%, Peer Reviewer = 37.50%, Institution = 12.50% E.g. Out of 40%, the 20% benefit should be passed on to researcher, 15 % benefit towards remuneration should be given to a reviewer and remaining 5% is to be retained by the institution.

We shall provide print version of 12 issues of any three journals [as per your requirement] out of our 38 journals worth $ 2376 USD.

**Other:**

**The individual Fellow and Associate designations accredited by Open Association of Research Society (US) credentials signify guarantees following achievements:**

- The professional accredited with Fellow honor, is entitled to various benefits viz. name, fame, honor, regular flow of income, secured bright future, social status etc.

- In addition to above, if one is single author, then entitled to 40% discount on publishing research paper and can get 10%discount if one is co-author or main author among group of authors.
- The Fellow can organize symposium/seminar/conference on behalf of Global Journals Incorporation (USA) and he/she can also attend the same organized by other institutes on behalf of Global Journals.
- The Fellow can become member of Editorial Board Member after completing 3yrs.
- The Fellow can earn 60% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.
- Fellow can also join as paid peer reviewer and earn 15% remuneration of author charges and can also get an opportunity to join as member of the Editorial Board of Global Journals Incorporation (USA)
- • This individual has learned the basic methods of applying those concepts and techniques to common challenging situations. This individual has further demonstrated an in–depth understanding of the application of suitable techniques to a particular area of research practice.

**Note :**

> "
>
> - In future, if the board feels the necessity to change any board member, the same can be done with the consent of the chairperson along with anyone board member without our approval.
> - In case, the chairperson needs to be replaced then consent of 2/3rd board members are required and they are also required to jointly pass the resolution copy of which should be sent to us. In such case, it will be compulsory to obtain our approval before replacement.
> - In case of "Difference of Opinion [if any]" among the Board members, our decision will be final and binding to everyone. "

# 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) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.**

**(II) Choose corresponding Journal.**

**(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.**

**(B) If you are using Internet Explorer, 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.**

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 Swis 721 Lt BT.
- 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 three 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 .

You must strictly follow above Author Guidelines before submitting your paper or else we will not at all be responsible for any corrections in future in any of the way.

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 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)

**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<br>Above 200 words | No specific data with ambiguous information<br>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 70116 58698 2 61427>

ISSN 9754350