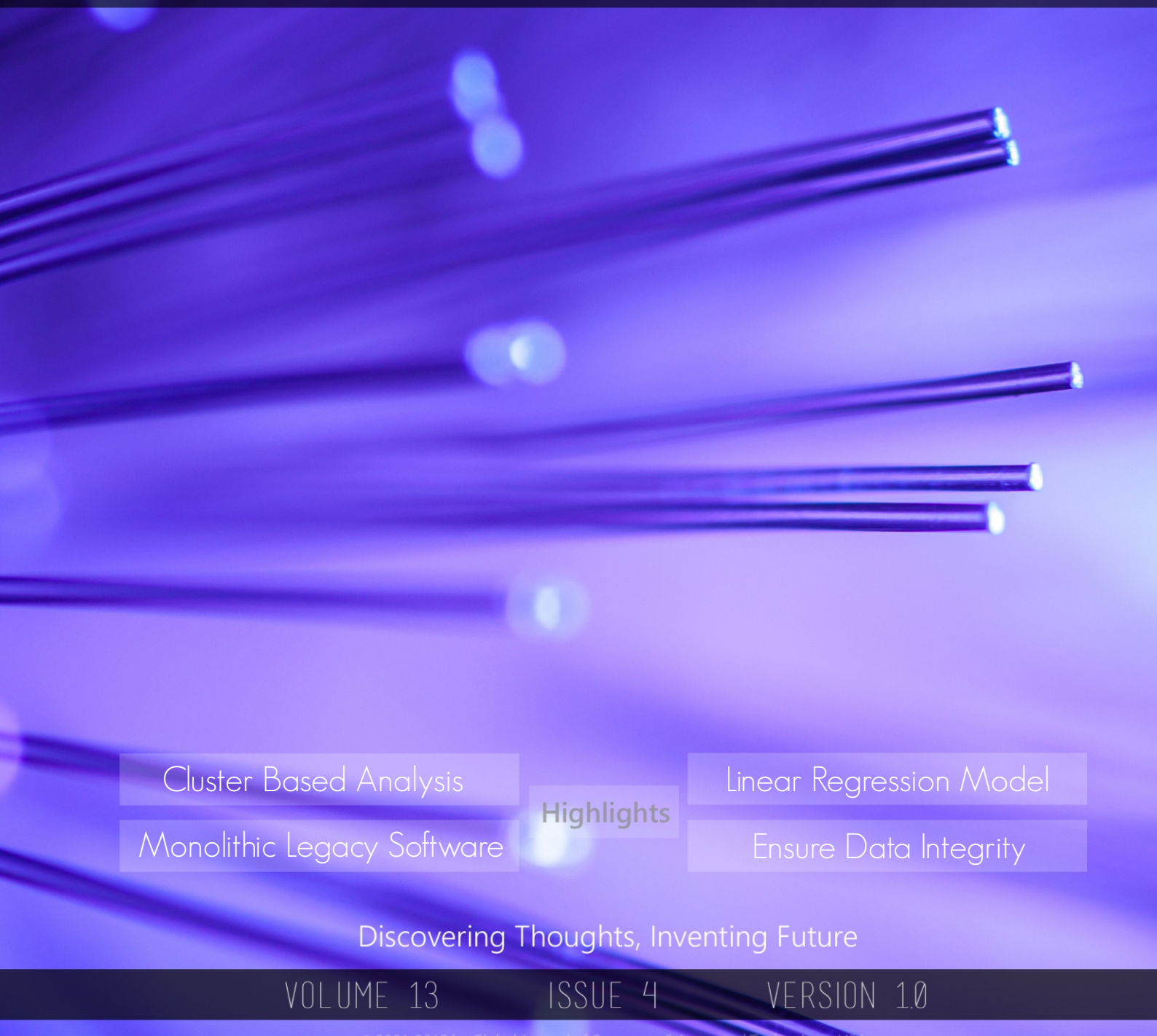
Online ISSN : 0975-4172
Print ISSN : 0975-4350

Global Journal

of Computer Science and Technology: C

Software & Data Engineering



Highlights

- Cluster Based Analysis
- Monolithic Legacy Software
- Linear Regression Model
- Ensure Data Integrity

Discovering Thoughts, Inventing Future

VOLUME 13 ISSUE 4 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 13 Issue 4 (Ver. 1.0)

Open Association of Research Society

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

Offset Typesetting

Global Association of Research, Marsh Road, Rainham, Essex, London RM13 8EU United Kingdom.

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

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

eContacts

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

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)
Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 250 USD (Color)

Editorial Board Members (HON.)

Dr. Bart Lambrecht
Director of Research in Accounting and FinanceProfessor of Finance
Lancaster University Management School
BA (Antwerp); MPhil, MA, PhD (Cambridge)

Dr. Carlos García Pont
Associate Professor of Marketing
IESE Business School, University of Navarra
Doctor of Philosophy (Management), Massachusetts Institute of Technology (MIT)
Master in Business Administration, IESE, University of Navarra
Degree in Industrial Engineering, Universitat Politècnica de Catalunya

Dr. Fotini Labropulu
Mathematics - Luther College
University of ReginaPh.D., M.Sc. in Mathematics
B.A. (Honors) in Mathematics
University of Windso

Dr. Lynn Lim
Reader in Business and Marketing
Roehampton University, London
BCom, PGDip, MBA (Distinction), PhD, FHEA

Dr. Mihaly Mezei
ASSOCIATE PROFESSOR
Department of Structural and Chemical Biology, Mount Sinai School of Medical Center
Ph.D., Etvs Lornd University
Postdoctoral Training,
New York University

Dr. Söhnke M. Bartram
Department of Accounting and FinanceLancaster University Management SchoolPh.D. (WHU Koblenz)
MBA/BBA (University of Saarbrücken)

Dr. Miguel Angel Ariño
Professor of Decision Sciences
IESE Business School
Barcelona, Spain (Universidad de Navarra)
CEIBS (China Europe International Business School).
Beijing, Shanghai and Shenzhen
Ph.D. in Mathematics
University of Barcelona
BA in Mathematics (Licenciatura)
University of Barcelona

Philip G. Moscoso
Technology and Operations Management
IESE Business School, University of Navarra
Ph.D in Industrial Engineering and Management, ETH Zurich
M.Sc. in Chemical Engineering, ETH Zurich

Dr. Sanjay Dixit, M.D.
Director, EP Laboratories, Philadelphia VA Medical Center
Cardiovascular Medicine - Cardiac Arrhythmia
Univ of Penn School of Medicine

Dr. Han-Xiang Deng
MD., Ph.D
Associate Professor and Research
Department Division of Neuromuscular Medicine
Davee Department of Neurology and Clinical NeuroscienceNorthwestern University
Feinberg School of Medicine

Dr. Pina C. Sanelli
Associate Professor of Public Health
Weill Cornell Medical College
Associate Attending Radiologist
NewYork-Presbyterian Hospital
MRI, MRA, CT, and CTA
Neuroradiology and Diagnostic Radiology
M.D., State University of New York at Buffalo,School of Medicine and Biomedical Sciences

Dr. Roberto Sanchez
Associate Professor
Department of Structural and Chemical Biology
Mount Sinai School of Medicine
Ph.D., The Rockefeller University

Dr. Wen-Yih Sun
Professor of Earth and Atmospheric SciencesPurdue University Director
National Center for Typhoon and Flooding Research, Taiwan
University Chair Professor
Department of Atmospheric Sciences, National Central University, Chung-Li, TaiwanUniversity Chair Professor
Institute of Environmental Engineering, National Chiao Tung University, Hsin-chu, Taiwan.Ph.D., MS The University of Chicago, Geophysical Sciences
BS National Taiwan University, Atmospheric Sciences
Associate Professor of Radiology

Dr. Michael R. Rudnick
M.D., FACP
Associate Professor of Medicine
Chief, Renal Electrolyte and Hypertension Division (PMC)
Penn Medicine, University of Pennsylvania
Presbyterian Medical Center, Philadelphia
Nephrology and Internal Medicine
Certified by the American Board of Internal Medicine

Dr. Bassey Benjamin Esu
B.Sc. Marketing; MBA Marketing; Ph.D Marketing
Lecturer, Department of Marketing, University of Calabar
Tourism Consultant, Cross River State Tourism Development Department
Co-ordinator , Sustainable Tourism Initiative, Calabar, Nigeria

Dr. Aziz M. Barbar, Ph.D.
IEEE Senior Member
Chairperson, Department of Computer Science
AUST - American University of Science & Technology
Alfred Naccash Avenue – Ashrafieh

Contents of the Volume

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Relevance of Agency Theory in Software Development

By Dipendra Ghimire

Abstract - In the Information technology field there has been lots of development. The development of software is increasing every day. IT professional are developing software for different business needs. Software is developed with internal IT professional as well as out sourcing. In many cases the outsourcing has been unsuccessful. In some cases the internal software development has also created some conflict. The relationship between the software developers and the project mangers is undesirable. This paper addresses these failed relationships and suggests a solution to a problem. The solution would be to diagnose the relationship from both sides. Secondly Agency theory can be implemented to resolve the conflict between the two.

Software development has been troublesome for many years. When actual result is compared to the desired and originally anticipated result, a large number of software project tend to run late, exceed the budget or may even be canceled. Now a days large number of organization are moving towards the software implementation either by outsourcing or through the software development department. When a contractor and runs develop software late, exceeds the budget there are often significant dispute between the development organization and the client who is funding the project. There may be the disputes that may lead to litigation for breach of conduct.

Agency theory is directed at the ubiquitous of agency relationship in which one party (principal) delegates work to another who performs the work. Agency theory is concerned with the resolving the problem between the principal and the client. It helps to resolve the two problems that arise in the agency relationship

GJCST-C Classification : *D.0*

Strictly as per the compliance and regulations of:

Relevance of Agency Theory in Software Development

Dipendra Ghimire

I. Introduction

In the Information technology field there has been lots of development. The development of software is increasing every day. IT professional are developing software for different business needs. Software is developed with internal IT professional as well as out sourcing. In many cases the outsourcing has been unsuccessful. In some cases the internal software development has also created some conflict. The relationship between the software developers and the project mangers is undesirable. This paper addresses these failed relationships and suggests a solution to a problem. The solution would be to diagnose the relationship from both sides. Secondly Agency theory can be implemented to resolve the conflict between the two.

Software development has been troublesome for many years. When actual result is compared to the desired and originally anticipated result, a large number of software project tend to run late, exceed the budget or may even be canceled. Now a days large number of organization are moving towards the software implementation either by outsourcing or through the software development department. When a contractor and runs develop software late, exceeds the budget there are often significant dispute between the development organization and the client who is funding the project. There may be the disputes that may lead to litigation for breach of conduct.

Agency theory is directed at the ubiquitous of agency relationship in which one party (principal) delegates work to another who performs the work. Agency theory is concerned with the resolving the problem between the principal and the client. It helps to resolve the two problems that arise in the agency relationship. The first problem is the conflict in the desire goal of principal and agent. And secondly it may be difficult for principal to verify what the agent is actually doing.

There are many conflicts in regard to the software development regarding the time, its final product and the cost increase in the process of development of the software. Agency theory would be helpful in maintaining the relationship between the principal and the client. In software is a technical field there are managers at the top level who may know how to use the software but many not be capable of understanding the depth of the software during the software specification process they might point out the requirements that are understand by them in a little different than the software developer. In this case the agency theory can play an important role in defining the requirements of the final product, the timeline needed and the cost required to develop the software.

Software development may contain issue that involves legal, economic, management and managerial and technological. Due to uncertainties about cost or technology the developer faces there is a risk of having to abandon the project at any level. The user many not be able to fully understand the system and might think from their point of view, Given the limited information, the management and the developer may make decision in their own interest. In this case we can relate the importance of the theory like agency theory.

Author : E-mail : dipendra_ghimire@yahoo.com

References Références Referencias

1. Krishna, S., Sahay, S., and Walsham, G. 2004. "Managing Cross-Cultural Issues in Global Software Outsourcing," Communications of the ACM (47:4), 01/04/2004, pp 62-66.
2. Mary S. Logan, (2000) "Using Agency Theory to Design Successful Outsourcing Relationships", International Journal of Logistics Management, The, Vol. 11 Iss: 2
3. Petter, S., Straub, D., and Rai, A. 2007. "Specifying Formative Constructs in Information Systems Research," MIS Quarterly (31:4), pp 623-656.
4. Raghu, T. S., Jayaraman, B., &Rao, H. R. (2004). Toward an integration of agent- and activity-centric approaches in organizational process modeling: Incorporating incentive mechanisms. Information Systems Research, 15(4), 316.
5. Robert D. Austin (2001) The Effects of Time Pressure on Quality in Software Development: An Agency Model.[ONLINE]
6. Rustagi, S., King, W.R., and Kirsch, L.J. 2008. "Predictors of Formal Control Usage in It Outsourcing Partnerships," Information Systems Research (19:2), pp 126-143.
7. Salger, F., and Engels, G. 2010. "Knowledge Transfer in Global Software Development: Leveraging Acceptance Test Case Specifications," ACM/IEEE 32nd International Conference on

Software Engineering, Cape Town, South Africa, pp. 211-214.
8. Sedera, D., and Gable, G.G. 2010. "Knowledge Management Competence for Enterprise System Success," The Journal of Strategic Information Systems (19:4), pp 296-306.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Cluster based Analysis and Consumption of Food Products in Targeted Public Distribution System

By Ms. P Shanmuga Priya & Dr. S Santhosh Baboo
D. G. Vaishnav College

Abstract - The Public Distribution System in India is 50 years old. At present it is being carried on as an anti-inflationary and antipoverty system. Tamil Nadu, the southernmost State in the country, is adopting the Universal Public Distribution System covering its entire population and supplying regularly rice, wheat, sugar, kerosene and other products like pulses, edible oil etc. The PDS is a centrally sponsored scheme that entitles beneficiaries to subsidized food grains every month. Several challenges have been identified in the implementation of PDS like (i) Targeting errors (ii) Large leakages or diversion (iii) The elimination of bogus cards and (iv) The problems in Fair Price Shops. This paper analyses and evaluates the problems and finds the possible solutions using the data mining techniques based on preprocessing and clustering. The K-means and K-harmonic means algorithms are combined to cluster the data based on the type of food commodities for rice and wheat.

Keywords : *K-means, K-harmonic means, PDS, cluster based PDS.*

GJCST-C Classification : D.2.8

Strictly as per the compliance and regulations of:

Cluster based Analysis and Consumption of Food Products in Targeted Public Distribution System

Ms. P Shanmuga Priya [α] & Dr. S Santhosh Baboo [σ]

Abstract - The Public Distribution System in India is 50 years old. At present it is being carried on as an anti-inflationary and antipoverty system. Tamil Nadu, the southernmost State in the country, is adopting the Universal Public Distribution System covering its entire population and supplying regularly rice, wheat, sugar, kerosene and other products like pulses, edible oil etc. The PDS is a centrally sponsored scheme that entitles beneficiaries to subsidized food grains every month. Several challenges have been identified in the implementation of PDS like (i) Targeting errors (ii) Large leakages or diversion (iii) The elimination of bogus cards and (iv) The problems in Fair Price Shops. This paper analyses and evaluates the problems and finds the possible solutions using the data mining techniques based on preprocessing and clustering. The K-means and K-harmonic means algorithms are combined to cluster the data based on the type of food commodities for rice and wheat.

Keywords : K-means, K-harmonic means, PDS, cluster based PDS.

I. Introduction

Data mining is defined as the discovery of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data[3]. It is also the extraction of interesting non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large databases. Preprocessing has been used to keep the data set ready for the process. Entity extraction has been used to automatically identify person, address and type of card. Filtering techniques has been used to filter the incomplete, noisy and inconsistent data.

a) Clustering

Clustering is the process of grouping data objects into similar classes. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is an important area of application for a variety of fields including data mining. It is the unsupervised classification of patterns (observations, data items or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines. This reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. There exist a large number of clustering algorithms in the literature. The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. The data mining clustering techniques are used to cluster the city ration card based on the type of cards. The preprocessed and clustered data are analyzed to evaluate the problems faced in the rationing system and the possible solutions are established.

In clustering the algorithms are applied to discover interesting data distributions in the underlying data space. The formation of clusters is based on the principle of maximizing similarity between patterns belonging to distinct clusters. Similarity or proximity is usually defined as distance function on pairs of patterns and based on the values of the features of these patterns [7] [8]. Many variety of clustering algorithms have been emerged recently. Different starting points and criteria usually lead to different taxonomies of clustering algorithms [9]. Some commonly used partitional clustering methods are K-Means (KM) [10], K-Harmonic Means (KHM) [11], Fuzzy C-Means (FCM) [12] [13], Spectral Clustering (SPC) [14] and several other methods.

b) Public Distribution System

The PDS is a centrally sponsored scheme that entitles beneficiaries to subsidized food grains every month [6]. Currently, beneficiaries are divided into the following groups: Below Poverty Line (BPL), Above Poverty Line and Antodaya Anna Yojana. The above mentioned challenges have been identified in the implementation of PDS. Some of them are as follows:

i. *Targeting Errors*

Separating beneficiaries of the PDS into three categories requires their classification and identification. Targeting mechanisms, however, have been prone to large inclusion and exclusion errors. Targeting error is where the targeted group of Below Poverty Line were denied ration cards. In 2009, it was estimated that about 61% of the eligible population was excluded from the BPL list while 25% of non-poor households were included in the BPL list.

Author α σ : Research Scholar, P.G. & Research Department of Computer Science, D. G. Vaishnav College, Arumbakkam, Chennai, Tamil Nadu, India – 600 106. E-mails : shanmusekar@gmail.com, Santhos2001@sify.com

ii. *Large leakages and diversion of subsidized foodgrain*

Foodgrain is procured by the centre and transported from the central to state godowns. Last mile delivery from state godowns to the Fair Price Shop (FPS) where beneficiaries can purchase grain with ration cards, is the responsibility of the state government. Large quantities of foodgrain are leaked and diverted into the open market during this supply chain.

iii. *Bogus Cards*

The bogus cards or false cards where a person owns duplicate cards and card in the name of non existing person.

iv. *Fair Price Shops*

The common problems faced by the public in the Fair Price Shops run by the state governments are underweight, poor quality and non availability of the essential commodities etc.

Targeted Public Distribution System (TPDS) is operated under the joint responsibility of the Central and the State/Union Territory (UT) Governments. Central Government is responsible for procurement, allotment and transportation of foodgrains upto the designated depots of the Food Corporation of India. Tamil Nadu Government is implementing Universal Public Distribution System (UPDS) and no exclusion is made based on the income criteria [1]. The state government has made the universal public distribution system 'poor friendly' by ordering rice at free of cost under public distribution system to all eligible card holders from 01.06.2011.

The public distribution system has been implemented through **33,222 fair price shops** comprising of **31,232** shops run by the **Cooperative Societies**, **1,394 shops** run by the **Tamil Nadu Civil Supplies Corporation** and **596 shops run by Women Self Help Groups**. Out of 33,222 fair price shops, **25,049** are **full time** shops and **8,173** are **part time** shops. Family cards are issued to the people of the State based on the option exercised by the individual family. The family cards are segregated as Rice Cards (rice with all other commodities), Antyodaya Anna Yojana Scheme cards, Sugar Cards (additional sugar in lieu of rice and all other commodities), and no ration commodity cards (cards for identification purposes). Besides, police personnel are issued with family cards in distinct colour. Transgender living in a house as a group is treated as a family and family cards are issued to them. As on 31.03.2012, **1,365 transgender family cards** have been issued [2]. This statistical information is taken from the website of Tamil Nadu civil supplies corporation.

The details of family cards in circulation in Tamil Nadu are as follows:

Sl. No.	Type of Card	Commodities entitled	No. of Cards
1.	Rice Cards	All Commodities	1,67,21,538
2.	Antoydaya Anna Yojana Cards	All Commodities	18,62,615
Total Rice Cards			**1,85,84,153**
3.	Sugar Cards	All Commodities except rice	10,76,552
4.	Police Cards	All Commodities	61,061
5.	No Commodity Cards	No Commodity	60,827
Total			**1,97,82,593**

II. Cluster Based Pds

In spite of the effort taken by the central and the state governments, the struggle faced by poor public, which is a usual scenario in almost every fair price shop is the main motivation for this research work. The main aim is to develop analytical data mining methods that can systematically address the composite problems assisting to

- Perform fault analysis to detect error patterns.
- Formulate an approach for error prevention and reduction.
- Categorize and analyse the common error patterns and minimise further occurrences of similar incidence.

The present research work proposes the use of an amalgamation of various data mining techniques to achieve the following objectives:

- To develop a data cleaning/filtering algorithm that cleans the dataset.
- To explore and enhance clustering algorithms to identify the error patterns from historical data.
- To explore and enhance classification algorithms to predict the future trends.

a) Methodology of Cluster based PDS

Pre-processing has been used to keep the data set ready for the process. Entity extraction has been used to automatically identify person, address, and type of card. Filtering techniques has been used to filter the incomplete, noisy and inconsistent data. The data mining clustering techniques are used to cluster the city ration card based on the type of cards. The preprocessed and clustered data are analysed to evaluate the problems faced in the rationing system and the possible solutions are established.

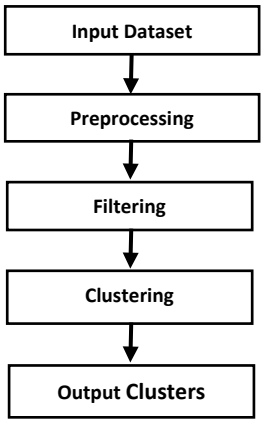


Figure 2.1 : The chart for cluster based PDS

i. *Preprocessing*

Most of the data collection techniques like survey studies, field experiments etc produce huge amount of information where the data are incomplete, noisy, inconsistent and contains missing values. Data preprocessing is a process that consists of data cleaning, data integration and data transformation which is usually processed by a computer program. It intends to reduce some noises, incomplete and inconsistent data. The results from preprocessing step can be used to proceed further by data mining algorithms.

ii. *Filtering*

Filtering is an initial step that is performed on the datasets in order to eliminate attributes that are unnecessary for the data mining process. This is an important task as some data mining tasks such as clustering (DB Scan, EM) and Association (Apriori) involving large amounts of data takes up large amount of time and memory. Filtering when not performed will slow down the application, to counter this problem the initial dataset instances which are populated by all the database attributes are filtered down to the most useful attributes depending on the applications requirements.

iii. *K-Means Clustering*

Given a set of objects, clustering is the process of class discovery, where the objects are grouped into clusters and the classes are unknown beforehand. Two clustering techniques, K-means and K-Harmonic means algorithms are considered for this purpose. The algorithm for k-means is given below [4].

$$V(D) = \sum_{i=1}^{k} \sum_{x_j = c_i} (x_j - u_i)^2$$

Where V is the variance, C_i is a cluster, u_j is its mean, D is the dataset of all points x_j. Partition the dataset into k clusters such that intracluster variance is minimized.

The primary part of the algorithm is the standard K-means algorithm. Lets us assume that current partition of the N p-dimensional point into k clusters, Compute the distances from each and every point to every cluster centroid and reassign. So for the simplest case of squared Euclidean distance, at every iteration, there is k computations of centroids, each one gets involved in p arithmetic means.

- k computations of centroids
  (Each of which involves p arithmetic means)
- $n^* k$ distance computations,
  (Each of which involves p sums of squared differences)
- n minimums over k distances

K-means algorithm often has hierarchical clustering using LINKAGE concepts. Hierarchical clustering needs n x n distance matrix, while K-means requires only the n x p data matrix, p is often much smaller than n.

iv. *K-Harmonic Means Clustering*

K-Harmonic Means Algorithm is a center-based, iterative algorithm that refines the clusters defined by K centers [7]. KHM takes the harmonic averages of the squared distance from a data point to all centers as its performance function. The harmonic average of K numbers is defined as the reciprocal of the arithmetic average of the reciprocals of the numbers in the set.

$$HA((a_i | i = 1,.....k)) = K\left[\sum_{i=1}^{k} \frac{1}{a_i}\right]^{-1}$$

The KHM algorithm starts with a set of initial positions of the centers, then distance is calculated by the function di, I =||xi-mi||, and then the new positions of the centers are calculated. This process is continued until the performance value stabilizes. Many experimental results show that KHM is essentially insensitive to the initialization. Even when the initialization is worse, it can also converge nicely, including the converge speed and clustering results.

III. Experimental Dataset

The dataset used in the present research work has been downloaded from the e-PDS portal Ministry of Consumer Affairs Food & Public Distribution, established to coordinate and integrate information resources created by the central government of India. The database is maintained by the central government of India. The dataset contains the allotted and off taken quantity of rice and wheat for all the states of India. This paper is concerned about only the state of Tamilnadu. The data are preprocessed, filtered and prepared for the clustering process with combined K- Means and K-Harmonic Means clustering techniques. The clustering process clusters the dataset based on the districts of Tamilnadu.

IV. Results and Discussions

Experiments were conducted to analyse the efficiency of the clustering methods for the allocated and off take quantity of food commodities like rice and wheat from the central government from 2001 to 2011 and the leakage or diversion of the commodities are calculated.

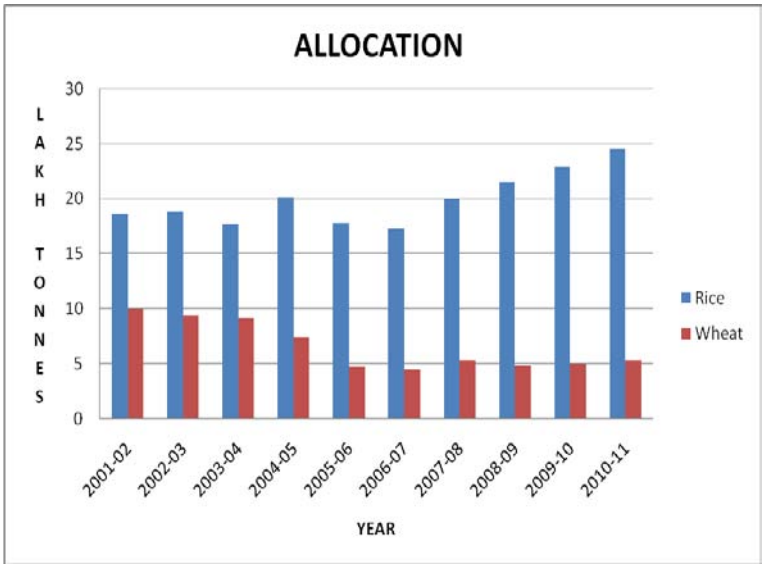


Figure 4.1 : Analysis of rice allotment

The graphs (Fig 4.1 & Fig 4.2) shows the alloted amount of rice and wheat from the central to the state government. The graph shows the amount of allocation and offtake of rice and wheat from the central government. There are ups and downs in the allotment from 2001 to 2005. The amount of rice alloted has been gradually increasing from 2006 to 2011. The offtake quantity of rice and wheat has been fluctuating from 2001 to 2011. In Fig 4.3 the diversion of wheat and rice has been calculated and the graph was plotted accordingly, which shows there is a steep increase in the diversion of rice in the year 2005-2006 and again a steep hike in 2010-2011 and the offtake of wheat is decreasing since 2001.

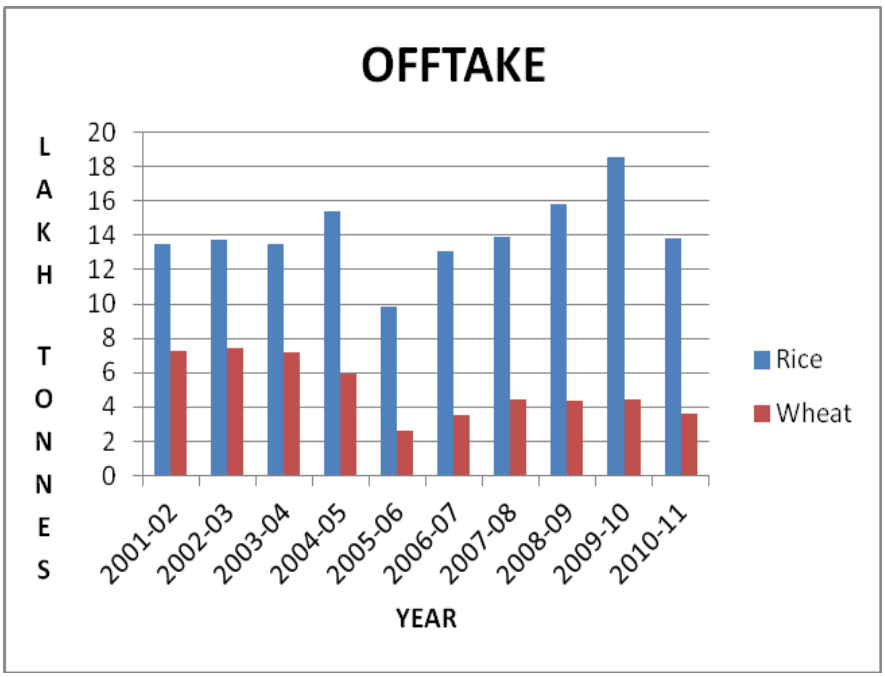


Figure 4.2 : Analysis of wheat allotment

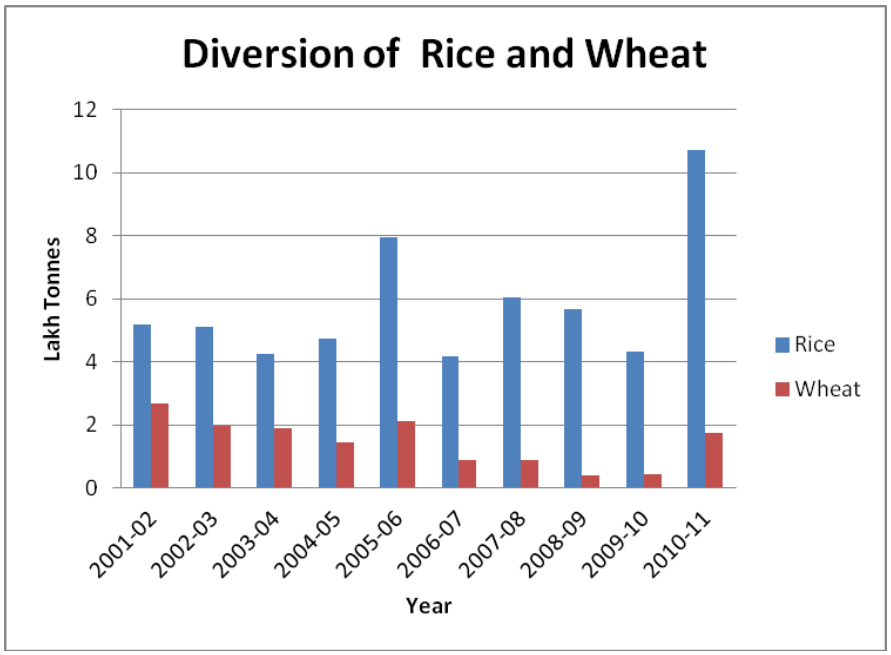


Figure 4.3 : Diversion of rice and wheat analysis

V. Conclusion and Future Work

Inspite of the steps taken from the central and the state government, eradication of the targeting problem can be established by the clustering techniques, where the different types of cards based on the food commodities of rice and wheat are clustered. The diversion of the food commodities can be tracked by the global positioning system installed in the transportation vehicles to the state godowns. This has been tested in the state of Tamilnadu and Himachal Pradesh and has been reported successively. The fair price shop management can be improved by sending SMS alert to the public about the food commodities availability and distribution on daily or weekly basis, so that the people can avoid the rush and repetitive visit to the fair price shop.

After Clustering the data can be classified to identify future fault that are emerging newly by using outlier detection on data. Experimental results prove that the algorithm is effective in terms of analysis speed, identifying common patterns and future prediction. The combined algorithm has promising value in the current changing scenario and can be used as an effective means by the Indian government PDS and Civil supplies corporation for fault detection and prevention.

References Références Referencias

1. Policy note on Food and Consumer Protection, 2012-2013, http://www.tn.gov.in/ policynotes/
2. Tamilnadu Civil Supplies Corporation, http://www.tncsc.tn.gov.in/html
3. Fayyad, U.M., & Uthurusamy, R. (2002) Evolving data mining into solutions for insights. Communications of the ACM, 45(8), 23-31.
4. Machine Learning & Computational Biology Research Group (2011), http://agbs.kyb.Tuebingen.mpg. de/ wikis/bg/2.pdf, (July 1, 2011).
5. Malathi. A, Enhanced Algorithms to Identify Change in Crime Patterns, International Journal of Combinatorial Optimization Problems and Informatics, Vol. 2, No.3, Sep-Dec, 2011, pp. 32-38, ISSN: 2007-1558.
6. The PRS Blog The official blogsite of PRS Legislative Research http://www.prsindia.org/theprsblog/
7. A.K. Jain, R.C. Dubes, "Algorithms for Clustering Data," Prentice-Hall, Englewood Cliffs, NJ, 1988.
8. J.T. Tou, R.C. Gonzalez, "Pattern Recognition Principles," Addision-Wesley, Reading 1974.
9. Chonghui GUO, Li PENG, "A Hybrid Clustering Algorithm Based on Dimensional Reduction and KHarmonic Means," IEEE 2008.
10. R C Dubes, A K Jain, "Algorithms for Clustering Data," Prentice Hall, 1988.
11. B Zhang, M C Hsu, Umeshwar Dayal, "K-Harmonic Means-A Data Clustering Algorithm," HPL- 1999-124, October, 1999. K. Elissa.
12. X B Gao, "Fuzzy Cluster Analysis and its Applications," Xidian University Press, 2004(in china).
13. Wu K L, Yang M S., "Alternative fuzzy c-means clustering algorithm," Pattern Recognition, 2002, 35(10): 2267-2278.

14. A Y Ng, M I Jordan, Y Weiss. "On spectral clustering: Analysis and an algorithm," In: Dietterich, T G Becker S, Ghahraman; Z. Advances in Neural Information Processing Systems 14, Cambridge: MITPress, 2002, 849-856.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Discriminative Gene Selection Employing Linear Regression Model

By Abid Hasan, Shaikh Jeeshan Kabeer, Md. Abdul Mottalib & Kamrul Hasan
University of Technology

Abstract - Microarray datasets enables the analysis of expression of thousands of genes across hundreds of samples. Usually classifiers do not perform well for large number of features (genes) as is the case of microarray datasets. That is why a small number of informative and discriminative features are always desirable for efficient classification. Many existing feature selection approaches have been proposed which attempts sample classification based on the analysis of gene expression values. In this paper a linear regression based feature selection algorithm for two class microarray datasets has been developed which divides the training dataset into two subtypes based on the class information. Using one of the classes as the base condition, a linear regression based model is developed. Using this regression model the divergence of each gene across the two classes are calculated and thus genes with higher divergence values are selected as important features from the second subtype of the training data. The classification performance of the proposed approach is evaluated with SVM, Random Forest and AdaBoost classifiers. Results show that the proposed approach provides better accuracy values compared to other existing approaches i.e. Relief F, CFS, decision tree based attribute selector and attribute selection using correlation analysis.

Keywords : *linear regression, feature selection, microarray dataset, classification.*

GJCST-C Classification : *D.2.2*

Strictly as per the compliance and regulations of:

Discriminative Gene Selection Employing Linear Regression Model

Abid Hasan [α], Shaikh Jeeshan Kabeer [σ], Md. Abdul Mottalib [ρ] & Kamrul Hasan [ω]

Abstract - Microarray datasets enables the analysis of expression of thousands of genes across hundreds of samples. Usually classifiers do not perform well for large number of features (genes) as is the case of microarray datasets. That is why a small number of informative and discriminative features are always desirable for efficient classification. Many existing feature selection approaches have been proposed which attempts sample classification based on the analysis of gene expression values. In this paper a linear regression based feature selection algorithm for two class microarray datasets has been developed which divides the training dataset into two subtypes based on the class information. Using one of the classes as the base condition, a linear regression based model is developed. Using this regression model the divergence of each gene across the two classes are calculated and thus genes with higher divergence values are selected as important features from the second subtype of the training data. The classification performance of the proposed approach is evaluated with SVM, Random Forest and AdaBoost classifiers. Results show that the proposed approach provides better accuracy values compared to other existing approaches i.e. Relief F, CFS, decision tree based attribute selector and attribute selection using correlation analysis.

Generalterms : algorithms, design, verification.
Keywords : linear regression, feature selection, microarray dataset, classification.

I. Introduction

The explosive growth and developments of microarray applications have enabled biologists and data mining engineers to study and observe thousands of gene expression data at the same time. Various attribute selection methodologies have been applied in the field of microarray data and in this particular case it is termed as gene selection as illustrated in [1]. Microarray data analysis has paved the way to cancer, tumor and other disease classification methods that can be used for subsequent diagnosis or prognosis. The problem of microarray data are many fold, firstly not all the data are relevant and often only a small portion of the data is related to the purpose of interest moreover noise and inconsistent data are prominent which hampers the search for the best genes for selection and classification [2]. However the major difficult aspect of microarray data is that the genes numbering in the thousands far outweighs the number of samples number in the lower hundreds if not less. This makes the task of building effective models particularly difficult and poses over fitting problems where the model does not perform well for novel patterns [3]. Thus feature selection methods being developed should be efficient in handling these issues.

Feature selection techniques can be generally divided into two broad categories depending on how the selection process interacts with classification model [4]. The first is the filter method where the importance of a feature is determined by scoring all the features based on their inherent attribute and retaining a portion of the features with higher scores while the low scoring features are removed as shown in many works including [5] and [6]. Filter methods are simple, fast and they do not require consultation with the classifier however the most obvious drawback is that it examines each feature individually and hence cannot harness the combined predictive power of features The second feature selection methodology is the wrapper model where a classification model is built by using a set of training set of features whose class labels are known and then the search for the optimal subset of features is done by repeatedly generating and evaluating possible feature using against the well known classifiers [7]. As the search for the solution is built into the classification process and as it considers the combinative predicting power of gene subsets the convergence time is higher the methods are usually complex.

Studies such as [8] and [9] have shown that the biological state of individuals is defined by their gene expression values. Therefore genes which have different expression profiles are more likely to properly identify biological states than genes having similar expression profiles. In this paper a linear regression model is proposed where one class of training dataset is considered as the base condition and generates the regression coefficients for each of the genes in the base class. Using the regression coefficients of the base condition a regression representation for the other class is generated and the difference in expression profiles between the genes of the base and non-base classes are measured. Genes with higher difference in expression profiles are given more importance and scoring of genes are generated. The base class serves as domain knowledge that is used to guide the search for discriminating genes in the dataset, [10] and [11] are works where domain knowledge was used to search for

Author α σ ρ ω : Department of CSE Islamic University of Technology, Gazipur 1704, Bangladesh. E-mails : aabid@iut-dhaka.edu, sjkabeer@iut-dhaka.edu, mottalib@iut-dhaka.edu, hasank@iut-dhaka.edu

the best features. The detailed procedure of the proposed method is provided in the following section. In the simulation and result analysis section it is seen that very high classification accuracy rates are achieved using only a very small number of the genes and the proposed method generated better results compared to other filtering approaches. The proposed approach has been applied on 6 microarray datasets and their effectiveness was determined by testing them in three different types of classifiers: Support Vector Machine (SVM), Random Forest and AdaBoost.

This paper is divided into 4 sections with section 1 giving an overview of the working domain and very brief introduction to the proposed approach. Section 2 elaborates the proposed approach in detail. Section 3 covers the simulation and result analysis part of the research where the proposed method is compared with Relief F, CFS, Chi-Squared value and Gain Ratio; it is seen that the proposed approach performs better then these existing methods. Section 4 provides the conclusion and provides scope for further research or development of this research work.

II. Proposed Approach

a) Theoretical Background

Linear regression is a statistical approach that can be used for predicting and forecasting. It has been traditionally used to model relationships between a set of explanatory variables $A=\{a_1,a_2,...,a_n\}$ and the output variable b_x. The idea is to derive a model using which the predictor or the output variable can be estimated using the explanatory variables [12]. In traditional feature selection applications the set of features are the input variables and the class labels are the output variables. Considering one feature a, the hypothesis function for this simple linear regression is

$$b_x = x_0 + x_1 a \tag{1}$$

where x_o and x_1 are the parameters and b_x is the predictor variable. The objective is to find the values of the parameters so that it best fits the data in the training set such that the features of unknown samples can be used for classification. x_j should be chosen such that $b_x(a)$ is as close to the training data (a,b) such that the following cost function is minimized

$$F(x_0,x_1)=\frac{1}{2m}\sum_{i=1}^{m}(b_x(a^i)-b^i)^2 \tag{2}$$

Here $F(x_0,x_1)$ is the cost function and m is the total number of samples in the training dataset. It is apparent that real world applications will require consideration of more than one feature and hence the hypothesis function will become

$$b_x(a) = x_0 + x_1 a_1 + x_2 a_{02} + x_3 a_3 + ... + x_n a_n \tag{3}$$

for convenience its assumed that $a_0 = 1$, therefore the feature vector A and parameter vector X becomes

$$A=\begin{bmatrix} a_0 \\ a_1 \\ a_3 \\ a_4 \\ . \\ . \\ . \\ a_n \end{bmatrix} \qquad X=\begin{bmatrix} x_0 \\ x_1 \\ x_3 \\ x_4 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

$b_x(a)$ can now be written as

$$b_X(a) = X^T A \tag{4}$$

X^T is the transpose of parameter vector and A is the vector of explanatory variables. So the corresponding cost function $F(X)$ which needs to be minimized for multiple variables, is the following:

$$F(x_0,x_1,x_2,...,x_n)=\frac{1}{2m}\sum_{i=1}^{m}(b_x(a^i)-b^i)^2 \tag{5}$$

Gradient descent is a very popular approach that has been used in many researches including linear regression. From earlier discussion it is clear that the idea is to minimize the cost function $F(X)$. Gradient descent algorithm helps find the parameter value which leads to the minimum cost. The representation of equation 5 in partial derivative term is

$$x_j := x_j + \alpha\frac{1}{m}\sum_{i=1}^{m}(b_x(a^i)-b^i)a_j^i \tag{6}$$

The algorithm starts with an arbitrary value x_j and keeps on changing by simultaneously updating x_j for $j=0,1,2,...,n$ until convergence for each of the x_j occurs.

b) Linear Regression on Microarray Dataset

Linear regression is a statistical approach that can be used for microarray datasets provides gene expression values for different samples. Using gene expression values to find out features and hence to classify novel samples is a common approach; however the application of linear regression to this task is a relatively fresh approach. In this proposed method the gene expression values of one class of samples of a two class microarray training dataset is used as the base class. Using this portion of dataset, a model is built which acts as the domain knowledge of the dataset.

Using this model the divergence in comparison to the gene expression values in the other class of the training dataset is measured. This tells us how much the latter deviates or diverges from the base class. From earlier discussions, equation 3 gives the expression for $b_x(a)$ where $A = \{a_1, a_2, ..., a_n\}$ are the features; for microarray applications, the genes are considered as features. As before $X = \{x_0, x_1, x_2, ..., x_n\}$ represents the parameters of the linear regression equation. In the proposed approach parameters X is being calculated by the cost function (equation 6) for each of the gene in the base class subtype of training dataset. Once we get the $n \times n$ parameter matrix X, it is applied to calculate $b'_x(i)$; the gene expression values in the non-base subclass of the training dataset. For each gene $b'_x(i)$ is calculated using all the gene data expect for its own hence

$$b'_x(1) = x_0 a_0 + x_2 a_2 + x_3 a_3 + ... + x_n a_n$$
$$b'_x(2) = x_0 a_0 + x_1 a_1 + x_3 a_3 + ... + x_n a_n$$
$$b'_x(3) = x_0 a_0 + x_1 a_1 + x_2 a_2 + ... + x_n a_n$$
$$\vdots$$
$$b'_x(n) = x_0 a_0 + x_1 a_1 + x_2 a_2 + ... + x_{n-1} a_{n-1}$$

These $b'_x(i)$ represent the statistical values of expression for each gene in the non-base subtype of the training dataset.

c) Proposed Algorithm

In our proposed method, basic idea of linear regression has been used. We have tried to predict a potential feature from one of the subtypes of microarray training datasets using the knowledge acquired from the other subtype of the same training dataset. At first the microarray dataset is divided into two segments test and training dataset in the similar way as most supervised learning algorithm does. One of the biggest problems of microarray data; redundancy has been handled by measuring the similarity in expression values of the genes in both types. We have eliminated those genes having similar expression values considering their ineffectiveness as important features for classification. Moreover, removing these genes gives the algorithm an efficient way of starting feature selection procedure. Training samples are then divided into two subtypes: S_1 and S_2 representing two subtypes of training data: base type and non-base type, built based on their class information. Next the parameter vector X for S_1 is generated using equation 6 and from the parameter vector X, $b'_x(a)$ is calculated for S_2. After the divergences and the differences are calculated, genes are sorted according to difference values in the descending order. From the sorted list of genes $N(N = 10,20,30,...,100)$ highest ranked genes are chosen and their classification accuracy is evaluated using different classifiers. Section 3 shows the detailed performance evaluation of the proposed approach and its superiority compared to other existing feature selection methods.

III. Materials and Methods

To find out how the proposed algorithm works, we have established the experiments using four different microarray datasets. We have compared our proposed feature selection algorithm with several other attribute selection procedures. Following sections describe a short description of microarray datasets and performance evaluations of the proposed method.

a) Datasets

The datasets are obtained from different authors. Datasets are converted into convenient way for this particular research.

The original prostate dataset was used in [13]. The dataset contains the 12,533 gene expression measurements of 102 samples. 50 of these 102 samples contain normal tissues not containing prostate tumor while 52 had prostate tumor.

Prostate cancer dataset was originally taken from dataset GSE2443 [14]. The dataset contains 12,627 gene expression values of 20 samples. Among them 10 samples contain androgen dependent tumor while other 10 contain androgen-independent tumor.

The lung cancer dataset contains two types of cancer: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of lung. Among 181 tissue samples, 31 of them had MPM and 150 of them had ADCA. Each of the samples was described by 12,533 gene expression value [15].

The colon dataset was used in [16]. The dataset contains 62 samples collected from colon cancer patients. 40 tumor biopsies are from tumors and 22 normal biopsies are from healthy parts of colon of same patients. The number of genes used in this expression is around 2000.

b) Performance Evaluation

The implementation of the proposed algorithm of feature selection was done on MATLAB and the performance evaluation of the selection set of features for classification was performed on publicly available weka tool [16]. We have used 10 fold cross-validation for SVM classifier. The random forest procedure was run with 10 trees and AdaBoost classifier uses 10 iteration and weighted threshold of 100.

i. *Results*

The classification accuracy of the features selected by proposed method and its comparison with

other method for different datasets is given in the following tables. *N* represents the number of features used by the classifier for classification.

Table 1 : Prostate dataset classification accuracy

Attribute selection method	Classifiers					
	SVM		Random Forest		AdaBoost	
	N	Acc (%)	*N*	Acc (%)	*N*	Acc (%)
ReliefF	100	80.33	150	81.97	100	88.52
CFS	17	67.21	17	59.02	17	67.21
Chi-Squared value	17	68.85	17	60.66	17	65.57
GainRatio Value	1190	83.61	1190	72.13	1190	85.24
Proposed	30	90.16	20	86.66	50	98.36

Table 2 : Prostate cancer dataset classification accuracy

Attribute selection method	Classifiers					
	SVM		Random Forest		AdaBoost	
	N	Acc (%)	*N*	Acc (%)	*N*	Acc (%)
ReliefF	200	58.33	150	66.67	100	50.00
CFS	44	58.33	44	25.00	44	33.33
Chi-Squared value	44	58.33	44	66.67	44	41.67
GainRatio Value	188	58.33	188	58.33	188	25.00
Proposed	20	91.67	20	75.00	10	83.33

Table 3 : Lung cancer dataset classification accuracy

Attribute selection method	Classifiers					
	SVM		Random forest		AdaBoost	
	N	Acc (%)	*N*	Acc (%)	*N*	Acc (%)
ReliefF	100	93.96	200	93.96	100	97.98
CFS	37	89.93	37	91.27	37	93.30
Chi-Squared value	37	89.93	37	92.62	37	94.63
GainRatio Value	705	91.27	705	95.30	705	97.31
Proposed	30	97.98	40	97.98	30	97.98

Table 4 : Colon dataset classification accuracy

Attribute selection method	Classifiers					
	SVM		Random forest		AdaBoost	
	N	Acc (%)	*N*	Acc (%)	*N*	Acc (%)
ReliefF	200	87.88	200	84.85	200	72.73
CFS	8	69.7	8	66.67	8	57.58
Chi-Squared value	8	81.82	8	69.70	8	63.64
GainRatio Value	62	81.82	62	66.67	62	69.69
Proposed	10	84.85	20	75.76	20	78.79

Another aspect of the proposed feature selection method is; we have not used any threshold for how many features for classification will be selected. Several different subsets of features have been used for classification and thus select the best subset based on its classifying ability. Figure 1 shows the error rate in classification by the classifier with any particular feature subset. For a particular feature selection *j* using a particular classifier *i,* the average error rate is calculated by $\frac{1}{3}\sum_{k} Accuracy(i, j, k)$ since we are evaluating the performance of a particular selector with different number of features.

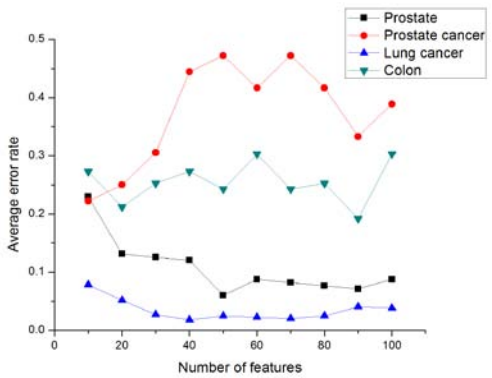


Figure 1 : Average error rate for different set of features

With the increase of the number of features, the error rate is decreased for most all the datasets. However, for prostate cancer dataset, although the error rate increases with first few subsets of features but at the end it too shows the same characteristics as the other microarray datasets shows.

ii. *Discussion*

We have proposed a new approach of feature selection using linear regression analysis. The algorithm works twofold. At the initial stage of the algorithm, we have eliminated redundant gene by measuring the similarity in expression values. Linear regression analysis then applied on one subtype (base type) of the training dataset to build the regression model. This model then applied on the other subtype (non-base type) of the training dataset to find out the divergence of the expression values of genes in that subtype. The more deviation shown by the gene, the more important it is considered as a feature. This way set of features selected for classification of the datasets.

Our main focus in this study is to classify accurately with less number of features. Table 1–4 shows the superiority of the classification accuracy by the features selected by the proposed method for different classifiers. Although, for colon dataset, classification accuracy by the features selected by ReliefF and CFS approach shows better result than the proposed method. However the result is still comparable and the number of feature selected by the proposed method is considerably fewer than the other method of attribute selection. Also Figure 1 summarizes the effect of different feature subsets for classification.

IV. Conclusion

Linear regression based feature selection shows promising results in classification of microarray datasets. The proposed approach might be applied on more microarray datasets and the results obtained might be used to improve some of the parameters of the proposed method. The results will also help to understand the performance of the proposed approach on a broader scale. The proposed approach can also be extended for multiclass approaches to be applied in other data mining domains. In the future Incorporation of other knowledge might help the proposed method to enhance the performance and significance of the result.

References Références Referencias

1. Kohavi, R and John, G.H. 1997. Wrappers for Feature Subset Selection, *Artificial Intelligence* Vol. 97 (1-2) (Dec. 1997), 273 - 324.
2. Iñaki Inza, Pedro Larrañaga, Rosa Blanco, Antonio J. Cerrolaza 2004. Filter versus wrapper gene selection approaches in DNA microarray domains, *Artificial Intelligence in Medicine* Vol 31 (2) (June 2004), 91–103.
3. Beatrice Duval and Jin-Kao Hao 2009, Advances in meta heuristics for gene selection and classification of microarray data, Briefings *in Bioinformatics* Vol. 11 (1) (July 2009), 127-141.
4. Yvan Saeys, Iñaki Inza, and Pedro Larrañaga 2007. A review of feature selection techniques in bioinformatics, *Bioinformatics* Vol.23 (19), 2507 -2517.
5. Hall M 1999. Correlation-based feature selection for machine learning. *PhD Thesis*. Department of Computer Science, Waikato University (1999).
6. Yu L, Liu H 2004. Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 2004; Vol. 5, 1205-1224.
7. Iñaki Inza, Basilio Sierra, Rosa Blanco, Pedro Larrañaga 2002. Gene Selection by Sequential Search Wrapper Approaches in Microarray Cancer Class Prediction, *Journal of Intelligent & Fuzzy Systems,* Vol. 12(1), 25-33.
8. Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, J.S. Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, Janos Demeter, Charles M. Perou, Per E. Lønning, Patrick O. Brown, Anne-Lise Børresen-Dale, and David Botstein, 2003. Repeated Observation of breast tumor subtypes in independent gene expression data sets, *PNAS,* Vol. 100 (14), 8418-8423.
9. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*. (Oct 15) Vol. 286 (5439), 531-7.
10. Ting Yu, Simeon J. Simoff and Donald Stokes 2007. Incorporating Prior Domain Knowledge into a Kernel Based Feature Selection Algorithm. *Lecture Notes in Computer Science,* Vol. 4426/2007, 1064-1071.
11. Ofir Barzilay, V.L. Brailovsky1999. On domain knowledge and feature selection using a support vector machine. *Pattern Recognition Letters,* Vol. 20, (5), (May 1999), 475–484.
12. X. Yan and X. Su. Linear Regression Analysis. *World Scientific,* 2009.
13. D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P.w. Kantoff, T. R. Golub, W. R. Seller, 2002. Gene expression correlated of clinical prostate cancer behavior. *Cancer cell*. Vol.1 (2). p.p. 203-9
14. C. J. Best, J. W. Gillespie, Y. Yi, G. V. Chandramouli, et al. 2005, Molecular alternations in primary prostate cancer after androgen ablation therapy. Clin cancer Res. Vol. 1(11), 6823-34.
15. G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. R. Ramaswamy, W.

G. Richards, D. J. Suqarbaker, R. Bueno, 2002. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and meothelioma. *Cancer Research.* Vol, 62. p.p. 4963-4967

16. U. Alon, N. Barakai, D. Notterman, K. Gish, S. Ybarra, D. Mack 1999, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *In Proceedings of National Academy of Science*. Vol. 96(12), 6745-6750
17. I. H. Witten and E. Frank 2005, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publisher.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

An Empirical Investigation for Understanding & Extraction of Services from Monolithic Legacy Software

By Asfa Praveen & Shamimul Qamar
Shri Venkateshwara University

Abstract - While working on modernization of large monolithic application; speed , synchronization and interaction with other components are the major concern for practical implementation of target system; as Service-Oriented Computing extends and covering many sections of monolithic legacy to web oriented development, these aspects becoming a new challenges to existing software engineering practices, the paper presents work which is undertaken for service orientation of monolithic legacy application including initial steps of service understanding, comprehension and extraction so that it can take a part in further migration activities to service oriented architecture platform. The work also shows that how several useful techniques can be applied to accomplish the result.

Keywords : *web services, ADT, SOA, clusters, comprehension.*

GJCST-C Classification : *D.2.11*

Strictly as per the compliance and regulations of:

An Empirical Investigation for Understanding & Extraction of Services from Monolithic Legacy Software

Asfa Praveen [α] & Shamimul Qamar [σ]

Abstract - While working on modernization of large monolithic application; speed , synchronization and interaction with other components are the major concern for practical implementation of target system; as Service-Oriented Computing extends and covering many sections of monolithic legacy to web oriented development, these aspects becoming a new challenges to existing software engineering practices, the paper presents work which is undertaken for service orientation of monolithic legacy application including initial steps of service understanding, comprehension and extraction so that it can take a part in further migration activities to service oriented architecture platform. The work also shows that how several useful techniques can be applied to accomplish the result.

Keywords : *web-services, ADT, SOA, clusters, comprehension.*

I. Introduction

A difficult and complex procedure for any maintenance project is the initial investigation which includes understanding of programs of software with its source code. This research is undertaken for service orientation of monolithic legacy software, till now many formal understanding and comprehension methods have been presented but conceptually and practically differ from one investigator to the other. Easy and quick monolithic legacy program understanding with fast comprehension is major concern of the work which plays a very important and crucial role in the planning, designing, feasibility study and cost estimation for services orientation projects of monolithic legacy [1]. The empirical examples/case studies have been presented to explain how the processes can be used to support better and improved comprehension in the program and incorporated services. The role of Ha-Slicer tool and web-mining techniques have been presented with application that appear to be reasonable for manual and automatically grouping extraction of services semantically similar in monolithic software and components [2]. The clusters of services understood, extracted by these processes represent an abstraction of the program source code based on a semantic similarities which should be incorporated further to high-level design of target system.

II. Problems for Understanding of Program

This section shows a sample program as presented in fig. 1, an analyzed program is depicted in the left hand side of the fig. 2, contains declarations, initializations and embedded print loop for each of three strings. The strings are considered as primitive data type for this illustration with no shared functionality for printing. To understand this program the library of program plan has to be considered, which has previously compiled knowledge for composition of program in this domain as shown in fig. 3; fig. 3 shows the library plan which contains the program plan [3] that contains the class string or abstract data type. The understanding of services and translation for source code with including singly abstract data type can be performed if the mapping can take place between original source and complied knowledge in the domain of services which is shown in fig. 2.

Author α : Ph.D. (Computer Sc.) Research Scholar, Faculty of Science & Technology, Shri Venkateshwara University, Gajraula, (U.P.), India.
E-mail : asfa_praveen@yahoo.com
Author σ : Professor of Electronics & Computer Engineering, Noida Institute of Engineering & Technology, Greater Noida, (U.P.), India.
E-mail : jsqamar@gmail.com

```
Class String {
                    char locStr [SIZE]
                    String( char* intStr )
                            for (int a=0; intStr [a]; a+ +)
                                    locStr[a] = intStr[a]; }
                    printString () {
                                    for ( int a=0; locStr[a]; a++);
                                    printf("%s", locStr[a];}}
```

Figure 1 : Sample Program

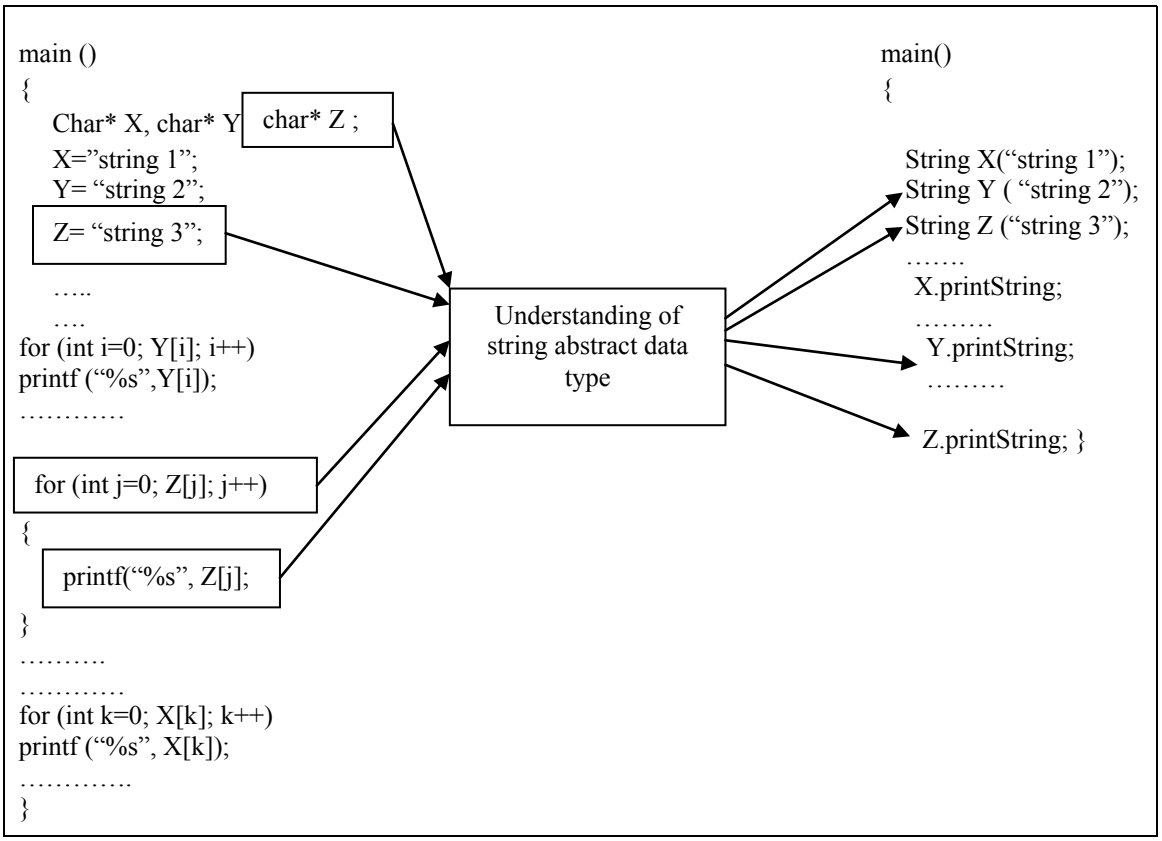


Figure 2 : Presentation of understating of mappings of C code abstract data type in view of object code [3]

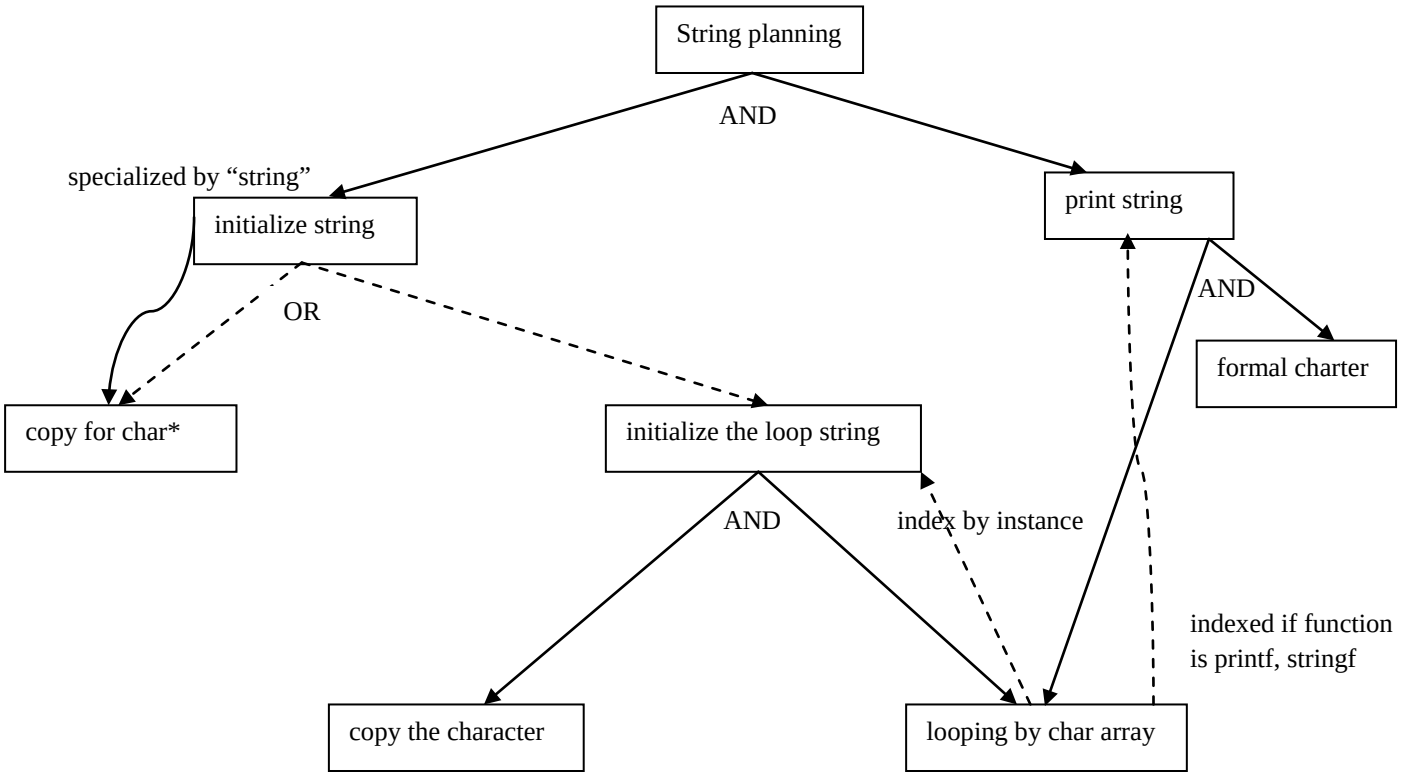


Figure 3 : String ADT within hierarchical plan program library [3]

Code in the left hand side of fig. 2 is given to explain the problem of understanding for program plan with the known context of as in the sting abstract data types. Identification of services and duplicate code may result in one time instance and incorporation of abstract data types. Abstract data types functionalities has been implemented same types in the right hand side of the fig. 2 in the object code the same concept has also been implemented in any other service oriented frameworks.

Suggested solution of the problems of understanding process is proposed in two phases, (1) investigate all instances of abstract data type in the source code with intention to convert them in services by abstract program features, (2) identify services plan blocks with program slices will relate to assure the hierarchical structure specified in the program plan based on knowledge. This can resolve the problem of knowledge management. Some useful advantages of identification are applied in the mapping of source code to the target service comprehension plan as when planning for replacing the source code by service oriented code resulting code will contain less code with same functionality and abstract data type and size for running, saving will be reduced, that will helpful for further implementation tasks. Mapping of the source code to the services is the main elementary blocks for the service oriented plan [4] and establishing plan library for either translation of code or identification of knowledge; it will reduce the bigger task of understanding.

III. Applications of Slicing, Haslicer Tools for Service Identification

Slicing based on Functional Dependence Graphs (FDG) contains five phases as illustrated in fig. 4 [5] the study conducted by Nuno et.al. The first phase parses the source code and originates the Abstract Syntax Tree instance t, which is followed by an abstraction procedure that extracts the useful information from *t* for constructing a FDG instance g with estimating the different types of nodes. Actual slicing is performed in the third phase, a slicing standard is composed here by a node of *t* and a specific slicing algorithm, the original FDG g is sliced, generating a sub-graph of g that is g'. The slicing takes place over the FDG to make the result which is always a sub-graph of the original graph.]

The fourth phase performs cutting AST t, based on the sliced graph g. At this point, each program entity that is not present in graph g*'*, is used to clip the correspondent syntactic entity in *t*, giving origin to a sub-tree t' of t. Finally, code reconstruction takes place, where the clipped tree *t'* is consumed to generate the sliced program by a reverse process of phase one.

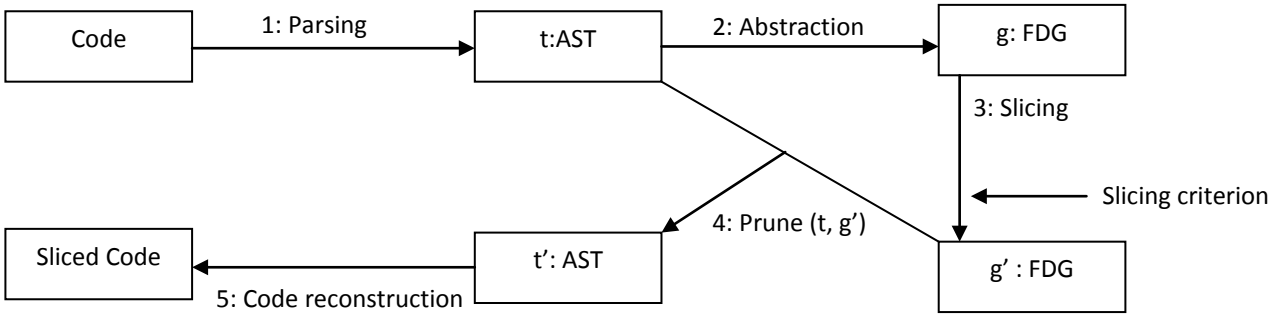


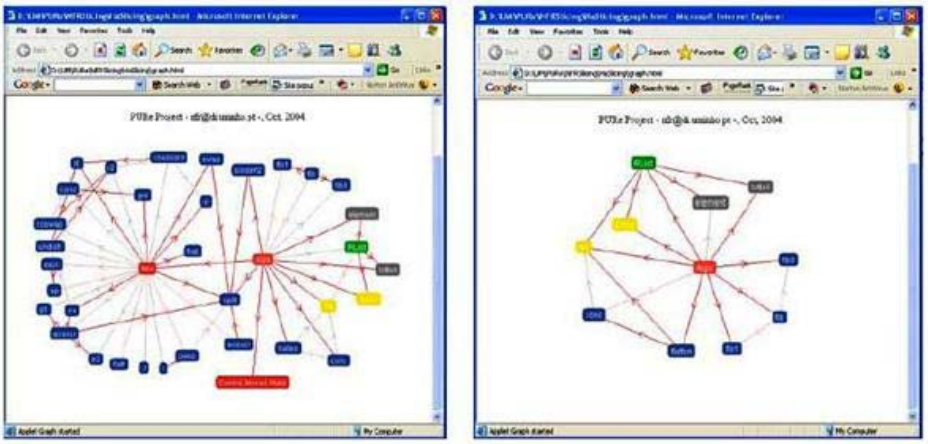
Figure 4 : Slicing Process [5]

Haslicer [6] is a sample tool for slicing for monolithic programs as sample here entirely written in Haskell language that will cover forward, backward and forward dependency slicing. The samples are sliced by implementing the slicing [7] and mention some other problems which are fundamentals to service component identification as (1) the definition of the extraction process from source code and (2) the incorporation of a visual interface by the generated FDG to support programmer interaction. It is accepting now only Haskell code [8] but can be plug-in for other monolithic code written in functional languages including purely functional language.

Fig.5 shows two snapshots of the sample working over a Haskell program [9]. Screenshot 5(a) shows the visualization of the entire FDG loaded in the tool. Notice that the differently colored nodes indicate different program entity types according to Table 1. Fig. 5(a) reproduces the sub-graph resulted from performing slice over one of the nodes of the graph from fig. 5(b). Once a slice has been computed, the user may retrieve the corresponding code. The whole process can be canceled or started again with different criteria.

Node Color	Node Type
	N_m
	N_f
	N_{dt}
	N_c
	N_d

Table 1 : FDG Edge Code



(a) (b)

Figure 5 : Slicing Process with HaSlicer [6]

There are basically two ways in which slicing techniques and the HaSlicer tool can be used in the process of service identification; either as a support procedure for manual identification or as a discovery procedure in which the whole system is searched for possible location of services. In this section both approaches are applied and briefly discussed, the first approach applied for manual service identification guided by analyzing and slicing some representation of the legacy code. In this context, the FDG seems to provide applicable representation model. Through its analysis, the program designer can identify all the dependencies between the code entities and look for certain architectural patterns or undesired dependencies in the graph. The process starts in a top-down way, looking for the top level functions that characterize the desired service, once these functions are found, forward dependency slicing is applied, starting from the corresponding FDG nodes. It produces a series of sliced code files, that have to be merged together in order to build the desired services. Forward dependency slice collects all the program entities in which each top level function requires to operate correctly. Thus, by merging all the forward dependency

slices corresponding to a particular service one gets the least derived program that implements it.

The second approach uses slicing mentioned in the beginning of this section under the name of component discovery relies on slicing techniques for the automatic isolation of possible components. As per experience this was found particularly useful at early stages of service identification. Such procedures, however, must be used carefully, since they may lead to the identification of both false positives and false negatives. This means that there might be good candidates for services which are not found as well as situations in which several possible services are identified which turn out to lack any practical or operational concern. To use an automatic service discovery procedure, one must first understand what to look for, since there is no universal way of stating which characteristics correspond to a possible software code component. Thus, process has to look for services by indirect means, which certainly include the identification of certain characteristics that components usually bear, but also some filtering criteria. Therefore a possible criterion for service discovery is based on the data types defined on the original code. The idea is to take each data type and isolate both data types and every program entity in the system that depends on it. Such an operation can be accomplished by performing a backward slicing starting from each data type node in the FDG.

Another identified feature for service-orientation task relates to the fact that interesting services typically present a high level of coupling and a high level of cohesion [10]. Coupling is a measure to estimate how mutually dependable two components services are, so it tries to measure how much a change in one service component affects other service components in a system whereas cohesion estimates how the functions as shown below [5] of a specific component which are internally related.

$$Coupling(G,f) \triangleq \sharp\{(x,y) \mid \exists x,y.\ yGx \wedge x \in f \wedge y \notin f\} \quad (1)$$

$$Cohesion(G,f) \triangleq \sharp\{(x,y) \mid \exists x,y.\ yGx \wedge x \in f \wedge y \in f\} \quad (2)$$

$$CCAnalysis(G) \triangleq \{(Coupling(G,f), Cohesion(G,f)) \mid \forall f \in \mathcal{P}F\} \quad (3)$$

In a service component with a low cohesion degree errors and undesirable behavior are difficult to detect. In practice if its functions are weakly related, errors may hide themselves in hardly ever used areas and remain unseen to testing for some time. The conjunction of these two measurement leads to discovery criteria, which uses the FDG to look for specific clusters of functions that is sets of strongly related functions, with reduced dependencies on any other program entity outside this set. Such function clusters cannot be identified by program slicing techniques, but the FDG is still very useful in determining this clusters. The HaSlicer tools compute the combined value where *G* is a FDG and *F* a set of functions under inspection. This is presented in study conducted by Nuno et.al. [5], which is depending on how easily or hardly service component discovery criteria are; then different acceptance limits for coupling and cohesion can be used. This will explain what clusters will be considered as location of prospective service components. Once such clusters are identified, the process continues by applying forward dependency slicing on every function in the cluster and merging the resultant code.

IV. Clustering Source Code Component's Services and Documents

Clustering for services of source code is based on semantic and structural information which is useful in the understanding and comprehension of monolithic software systems, on the other hand clustering can be used to support in re-modularization of systems and the identification of services from abstract data types [11]. If the program is to be reengineered for a service-oriented platform from a monolithic program this type of clustering would be very useful. The purpose is to decrease the quantity of source code when an engineer wants to observe at once and guess about possible relationships with the system not obvious from the current organization's documentation.

The method used here focuses on using reports generated by conceptual approaches, for this case a vector represented by latent semantic indexing [12] is generated to compare services and classify them into clusters of semantically similar concepts and for huge program it can be partitioned into a group of only source code documents by which the features for each document are prepared. Program documents are divided semantically based on similarities for connecting other documents to cluster the source code. For this purpose there are many applied clustering algorithms: construction, optimization, hierarchical and graph theoretical algorithms. There are also several other algorithms that use notion of hybrid concept applied for different classification for specific problems. The framework here proposes graph theoretic approach although numbers of other types of clustering algorithms have been used to cluster software program.

To cluster the documents minimal spanning tree algorithm [13] can be is used if the document is similar to at least some documents in the cluster then it is added to cluster, this can give an opportunity to group together the documents of similar type. The measurement of similarity is calculated by the cosine of the two vector representations of the source code documents. The similarity values has [-1, 1] for a domain, with the value 1 being closely similar. Non required symbols can be removed by using simple parsing of the source code that can break the source into the proper way. Comment delimiters and syntactical reserved words are removed; they had to add little or no semantic knowledge for problem domain [14]. On the other hand information retrieval process will analyses such confusing words such as semi-colons with a completely non-selective characteristic between source code and service components. So the variation with this characteristic is very little like zero thus; if two components have a semi-colon then not sure about their similarity. For the uses of latent semantic indexing on natural language perspectives a paragraph or code section is used for document because sentences are to be small and chapters too large. Source code that has similar concepts are: function, structure, module, file, class, etc. Observably the statement granularity is very low and file containing many functions can be too big.

V. Applying Web-Mining Techniques to Understand Services

So many techniques have been developed in Web-mining for successfully analyze the structure of web-services [16]. These techniques undertake the internet based web as a large graph which is based on hyperlink structure to identify the intended web pages. This section presents the study shows the application of web mining techniques, how to apply them to trace and understand classes and web services. HITS web-mining algorithm [15] is suggested to identify hubs and authorities for the web services. The HITS algorithm can be combined with the compressed call graph. The classes which are related with excellent "hubs" in the compressed call graph are good candidates for introduction of aspects.

a) Identifying Hubs and Authority in Big Web-Graphs

The concepts of "hub" and "authority" are introduced by HITS web-mining algorithm [15], hubs are pages that refer to pages containing information rather than being enlightening themselves, for examples web directories, lists of personal pages etc. and a page is called an authority if it has useful information. Thus, a web-page is a good hub if it is providing useful information. A page can be called as good authority if it is used by many good hubs. The HITS algorithm is based on this relation between hubs and authorities. This example considers the web-graph shown in fig. 6.

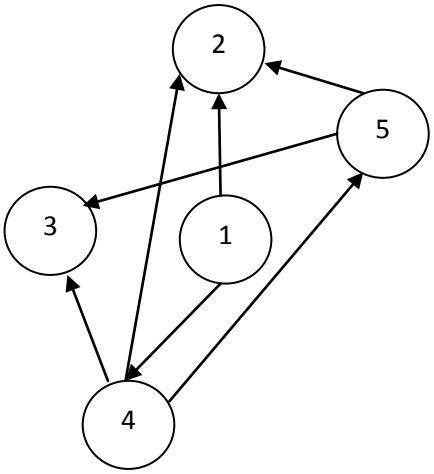


Figure 6 : Example Web-Graph

In this graph, 2 and 3 are good authorities, and 4 and 5 are good hubs, and 1 is a less good hub. The authority of 2 is larger than the authority of 3, because the only in-links that do not have in common are 1 2 and 2 3, and 1 is a better hub than 2. 4 and 5 are better hubs than 1, as they point to better authorities. The HITS algorithm works as follows: Every page i get assigned to it two numbers; ai denotes the authority of the page, while hi denotes the hubiness. Let i j denote that there is a hyperlink from page i to page j. The recursive relation between authority and hubiness is confined by the following formula.

$$h_i = \sum_{i \to j} a_j \quad (1)$$

$$a_j = \sum_{i \to j} h_i \quad (2)$$

The HITS algorithm starts with initializing all h's and a's to 1, and repeatedly updates the values for all pages, using the formula (1) and (2). If after each update the values are normalized, this process converges to stable sets of authority and hub weights [15]. It is also possible to add weights to the edges in the graph. Adding weights to the graph can be interesting to capture the fact that some edges are more important than others. This extension only requires a small modification to the update rules. Let w [i, j] be the weight of the edge from page i to page j. The update rules become

$$h_i = \Sigma_{i \to j}\, \omega[i, j].a_j$$

and

$$a_j = \Sigma_{i \to}\, \omega[i, j].h_i$$

Example: For given graph, the hub and authority weights to the following values:

$h_1 = 64$ $a_1 = 0$
$h_2 = 48$ $a_2 = 100$
$h_3 = 0$ $a_3 = 94$
$h_4 = 100$ $a_4 = 24$
$h_5 = 100$ $a_5 = 0$

In the context of web-mining, the identification of hubs and authorities by the HITS algorithm has turned out to be very useful. Because HITS only uses the links between web-pages then can be used in services [15].

VI. Service Extraction Process

The initial three steps of the service extraction process [18] as shown in fig. 7 represent the candidate service identification phase; candidate service identification is a challenging job, so a step-wise identification approach is designed. (a) Initially, the research finds how to utilize architectural reconstruction and source code visualization techniques. This step facilitates the proper understanding of code and to obtain structural properties of the source code [17]. The source code visualization technique presented by Geet et.al.[19] appears to be a good starting point. (b) The next step is to identify the design patterns [Gamma, 1995], one of the largely studied and applied techniques in context of reverse engineering. This is the extension for design pattern detection and its applicability in legacy to service oriented migration. (c) In the last step, linguistic analysis techniques are used [20] and concept analysis is used to find appropriate concepts that have been applied in the source code. The service extraction is performed after the application of concept slicing technique which can be further applied to extract the complete code generating the identified functionalities; it is fairly used to extract from source code various useful features for program comprehension [20]. It can independently extract from source code with the help of code query method. [21], that helps to extract abstract data type and common concern features, this extraction maps source code to service composition, the language features then supports the fixing composition related issues, in order to build new services. The main advantage of extraction is to generate services by component effective approaches then compose them to achieve target system implementation. With all these steps, process has achieved a simplified identification of candidate services in the monolithic code.

a) Service Understanding and Extraction Guidelines

There are following guidelines [22] for extraction and understanding of services;

i. *Realistic Representation*

Any service understood from the code must present real functional state in the program. This is the most important criteria if this is having any conflict with any other then this should be given priority because initial investigation goes through the program which is more trustworthy asset than a documentation.

ii. *Multiple representations for different abstraction of hierarchy*

Different development teams require different representations of services, for example, a programmer would like to have web services represented as code segments, while developer may require different types of forms as decision table, tree or chart to get the logical structure; so it must be represented in a hierarchy oriented way. Service understanding is more complex if they are in various constraints based perspectives, such as legal, marketing and technology. It is hardly tuff task to trace services without some form of abstractions or decomposition of the program.

iii. *Domain Oriented Policies*

Services expressed in the domain specific environment are far better because it connects with domain concept and propagates path for easy application. Then many tools can be applied for identification of logic of algorithm, data structures and other program entities.

iv. *Human-Assisted Automation*

As monolithic programs are huge having a lot of difficulties so if not impossible it is devised to use semi-automatic tool for service understanding. The software maintainers prefer to have an interactive tool that allow them to extract services, simplify their representations, and provide linkage to the code, rather than providing a black-box tool that generate services code automatically.

v. *Maintenance Tool*

Understood services will be useful in other software reengineering activities. Understanding should be managed together with the monolithic software using the same tool as the mapping from any service to its corresponding code segments. This capability will permit the software engineers to focus on only those segments and functions of the software that is relevant to a particular service type.

VII. Conclusion

This paper focused on program understanding and extraction for service orientation of monolithic code and presented various applied methods, techniques as clustering, web-mining, slicing, reverse engineering, and some more used tools that facilitates the automated process as Ha-Slicer with their previous applications, also presented the results and procedures, which was started with the understanding of code and finally concluded on the extraction of services.

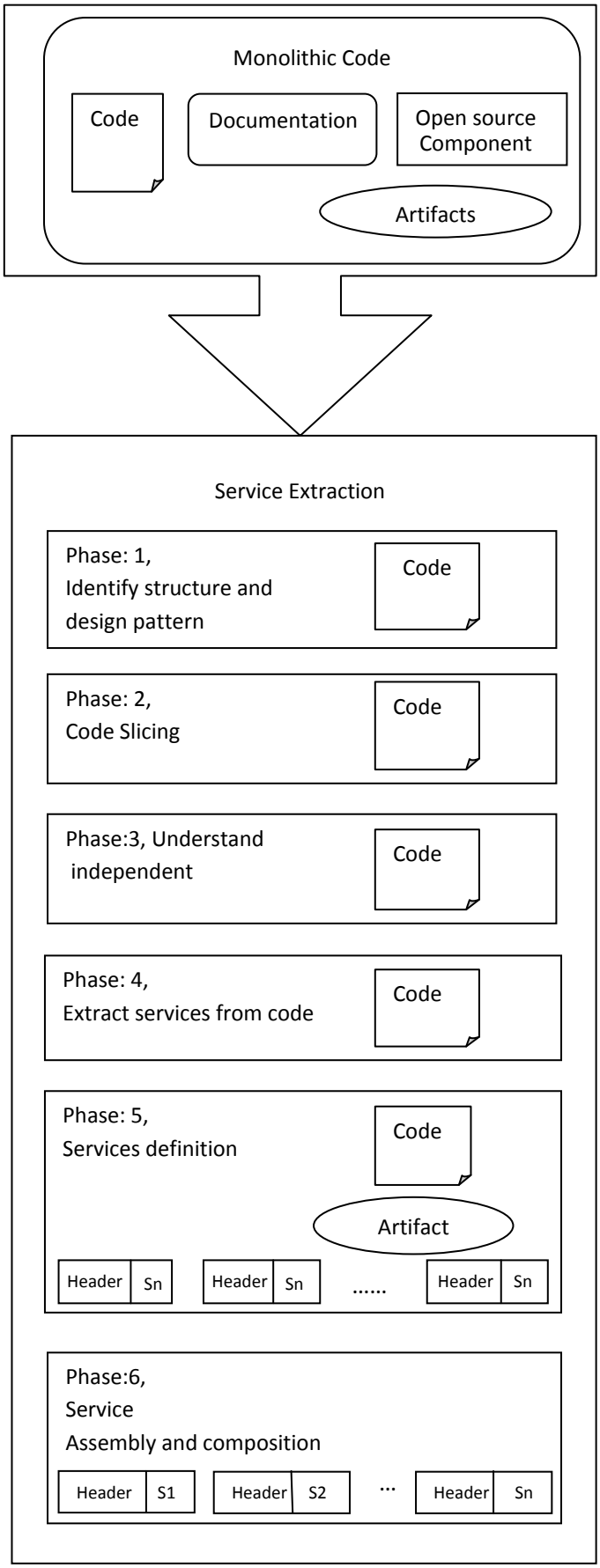


Figure 7 : Service Extraction Process

References Références Referencias

1. http://www.podcast.com/Technology/l-9445.htm
2. Shahanawaj Ahamad, "Web Centric Evolution of Legacy System", International Journal of Electrical and Computer Science, Vol: 10, No:1, pp: 19-24, 2010.
3. Steven Woods and Qiang Yang, "Program Understanding as Constraint Satisfaction", IEEE, Department of Computer Science, University of Waterloo, Canada, 1995.
4. http://technet.microsoft.com/en-us/magazine/ee677579.aspx
5. Nuno F. Rodrigues, Lu´ıs S. Barbosa, "Component Identification Through Program Slicing", in Electronic Notes in Theoretical Computer Science, Science Direct, Elsevier, pp: 291-304, 2006.
6. Nuno Miguel et.al, "Slicing Techniques Applied to Architectural Analysis of Legacy Software", report of Departamento de Informatica Escola de Engenharia, Universidade do Minho, 2008.
7. N. Rodrigues, "A basis for slicing functional programs". Technical report, PURe Project Report, DICCTC, U. Minho, 2005.
8. http://www.haskell.org/haskellwiki/Haskell
9. http://en.wikipedia.org/wiki/Haskell_(programming_language)
10. http://www.shu.ac.uk/softeng/extrabits/modularity/modularity%20-%20new%20version.doc
11. Adrian Kuhn et. al., "Semantic clustering: Identifying topics in source code", Information and Software Technology, Elsevier, pp: 230–243, 2007.
12. T. K. Landauer and S.T. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge", Psychological Review, vol. 104, no. 2, pp. 211-240, 1997.
13. J. B. Kruskal, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem", Proc. Amer. Math. Soc., vol.7, no. 1, pp. 48-50, 1956.
14. Jonathan I. Maletic et.al, Supporting program comprehension using semantic and structural information, Proceeding ICSE '01 Proceedings of the 23rd International Conference on Software Engineering Pp: 103-112, 2001.
15. J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, 46(5): 604–632, 1999.
16. D. Gibson, J. M. Kleinberg, and P. Raghavan. "Inferring web communities from link topology", In UK Conference on Hypertext, pages 225–234, 1998.
17. http://www.erikvanveenendaal.nl/NL/files/e-book%20PRISMA.pdf
18. http://www.docstoc.com/docs/94425343/Journal-of-Computer-Science-and-Research-Vol-9-No-8-August-2011

19. J. Van Geet and S. Demeyer, "Lightweight visualisations of cobol code for supporting migration to SOA," in 3rd International ERCIM Symposium on Software Evolution, October, 2007.
20. N. och Dag, B. Regnell, V. Gervasi, and S. Brinkkemper, "A linguistic engineering approach to large-scale requirements management," Software, IEEE, vol. 22, no. 1, pp. 32–39, 2005.
21. S. R. Tilley, D. B. Smith, and S. Paul, "Towards a framework for program understanding," in 4th International Workshop on Program Comprehension (WPC'96), 1996, pp. 19–28.
22. http://www.infosys.com/infosys-labs/publications/Documents/knowledge-engineering-management.pdf

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Dynamic vs Static Term-Expansion using Semantic Resources in Information Retrieval

By Ramakrishna kolikipogu & Padmaja Rani B
Sridevi Women's Engineering College

Abstract - Information Retrieval in a Telugu language is upcoming area of research. Telugu is one of the recognized Indian languages. We present a novel approach in reformulating item terms at the time of crawling and indexing. The idea is not new, but use of synset and other lexical resources in Indian languages context has limitations due to unavailability of language resources. We prepared a synset for 1,43,001 root words out of 4,83,670 unique words from training corpus of 3500 documents during indexing. Index time document expansion gave improved recall ratio, when compared to base line approach i.e. simple information retrieval without term expansion at both the ends. We studied the effect of query terms expansion at search time using synset and compared with simple information retrieval process without expansion, recall is greatly affected and improved. We further extended this work by expanding terms in two sides and plotted results, which resemble recall growth. Surprisingly all expansions are showing improvement in recall and little fall in precision. We argue that expansion of terms at any level may cause inverse effect on precision. Necessary care is required while expanding documents or queries with help of language resources like Synset, WordNet and other resources.

Keywords : *information retrieval, query expansion, semantics, indexing, document expansion, information retrieval in indian languages.*

GJCST-C Classification : *H.3.3*

Strictly as per the compliance and regulations of:

Dynamic vs Static Term-Expansion using Semantic Resources in Information Retrieval

Ramakrishna kolikipogu [α] & Padmaja Rani B [σ]

Abstract - Information Retrieval in a Telugu language is upcoming area of research. Telugu is one of the recognized Indian languages. We present a novel approach in reformulating item terms at the time of crawling and indexing. The idea is not new, but use of synset and other lexical resources in Indian languages context has limitations due to unavailability of language resources. We prepared a synset for 1,43,001 root words out of 4,83,670 unique words from training corpus of 3500 documents during indexing. Index time document expansion gave improved recall ratio, when compared to base line approach i.e. simple information retrieval without term expansion at both the ends. We studied the effect of query terms expansion at search time using synset and compared with simple information retrieval process without expansion, recall is greatly affected and improved. We further extended this work by expanding terms in two sides and plotted results, which resemble recall growth. Surprisingly all expansions are showing improvement in recall and little fall in precision. We argue that expansion of terms at any level may cause inverse effect on precision. Necessary care is required while expanding documents or queries with help of language resources like Synset, WordNet and other resources. Expansion techniques sometimes lead to poor performance and may miss the concept too. This increases overhead on naïve users to decide relevancy of outcome. Exhaustivity must be low to control adverse effect of precision and balance the recall as well. The same approaches are adapted to huge document collection from Wiki-Telugu and studied the effect.

Keywords : information retrieval, query expansion, semantics, indexing, document expansion, information retrieval in indian languages.

1. INTRODUCTION

Information Retrieval in local languages is getting more popularity in developing countries like India. Use of Internet and other Information Accessing Systems plays major role in Education, Medical, Business, Agriculture and other significant domains. Information Retrieval Systems in Local Languages that are getting popular among the Netigens, who prefer to access their information needs in their mother tongue language. Availability of digital documents in native languages creates interest to the user to access the information by typing query in local languages. India is multilingual country and people across the country speak more than 400 languages, but all the languages are not recognized due to lack of scripts and rules. The government of India has given "languages of the 8th Schedule" official status for 22 languages. Telugu is one of the recognized languages of India. Processing of Telugu digital items is more difficult when compared to European languages and other Indian Languages.

Building efficient Information Retrieval System for Telugu is a challenging task due to richness in Morphology and conflational features of the language. In this paper we studied the effect of Document Expansion and Query Reformulation Techniques with the help of synset lexical resource. Naïve users prefer to give one time query and expect adequate results in the first glance. In general lot of fuzziness involved in user query and it is difficult to match the relevant items. There is a necessity to reduce the vocabulary mismatch between naïve user query and repository prepared by domain experts. This paper is an attempt to study the impact of terms expansion during Indexing and searching on a sample Telugu text corpus.

When we expand the terms during indexing and supplied normal un-expanded queries to the Information Retrieval System, we observed that, there is a great fall in precision. Based on the synset length for each term the recall is positively affected. We then tested the same system by expanding query terms and keeping unexpanded items as source. Surprisingly the effect is similar and found improvement in recall and negative effect on precision. Expansion of terms either from document or query, the precision and recall are inversely proportional in growth rate. Main Objective of Information Retrieval is to retrieve relevant information from huge repository and preset top ranked items to the end user by reducing overhead in terms of time. Naïve users may not give strong queries to represent the concept in which, they expected to retrieve. Terminology of user for writing a query is always simple and vague; it may not resemble the concept of an item to be retrieved. We mean naïve user's vocabulary may generally drawn from day to day usage language, where as resources are drafted content in expertise vocabulary. Sometimes user may fail to use expertise vocabulary to write queries and represent the concept.

Most of the systems work on syntactic base; it requires exact matching of terms from query to document. Syntactic patterns are words in text mining. Word mismatch is a severe problem in Information [1].

Author α : Associate Professor & Head of Information Technology Department, Sridevi Women's Engineering College, Hyderabad, India. E-mail : krkrishna.csit@gmail.com

Author σ : Professor & Head of Computer Science and Engineering Department, JNTUH College of Engineering, JNTUH University Hyderabad, India. E-mail : Padmaja_jntuh@yahoo.co.in

In this paper we present various term selection methods for query reformulation and item expansion with implementation along with results as listed:

1) Simple IR System using statistical Indexing with nterms length query.
2) Query Reformulation at runtime using term expansion with synset in Pseudo Relevance Feedback (PRF) Approach.
3) Item Reformulation based on query terms using PRF approach.
4) Query Expansion and item expansion using synset with blind retrieval approach.

These approaches were discussed in Chapter 3 and Results are given in Chapter 5.

II. Related Work

Information is growing in an exponential manner on World Wide Web, the problem of finding useful information and knowledge from abundant source becomes one of the most important topics in information retrieval and storage [4], [5]. Information retrieval support systems are being developed in supporting users to find necessary information and knowledge [3]. Information Retrieval System is a multidiscipline area of research, which involves text processing, speech processing, image processing, video processing and other mode of information processing. Retrieval of any kind of information mainly aims at satisfying end user to his query. Usually naïve users search with text query by limited vocabulary. Representation of source in order to facilitate matching against user query plays major role and having equal importance with query structure. In this paper we limited to text documents as resource to retrieve for the given query. Many of the documents retrieved for general queries are totally irrelevant to the subject of user interest, due to insufficient keywords supplied in the search [6]. Sometimes the words entered by user may not express the interest of the user. Vocabulary of users may far from the expert's terminology in documents and it is difficult to match the same. The word mismatch can be solved by rewriting queries with new terms called as query expansion [7]. Our objective in this paper is to select a suitable term for expansion and to improve precision and recall as well. Level of query expansion varies from model to model. Expansion Terms can be selected in many ways 1) Suggested terms are provided to select by user and expand the query without missing concept. This is more accurate way of term expansion called manual expansion, but it requires knowledge to judge the term relevance, which increases overhead on user. Naïve users are not familiar in writing queries; hence the word miss match comes into the picture. User can not be given burden to use retrieval system, that's why automatic query expansions techniques are regular practice in IR Systems. Relevance Feedback [8] method considers user selection out of retrieved as relevant and reformulate the query to repeat the search by adjusting weights of initial query terms. Users who are familiar with query expansion takes maximum benefit of query reformulation [9] with relevance feedback. In expertise user will better serve with Pseudo Relevance Feedback called Automatic Query Expansion [10].

Information retrieval using Language models are used to improve relevance of a query outcome by document set feedback [11]. Cluster Feedback (CFB) is another way of term selection to find more similar terms by clusters. If relevant clusters are identified, then combining them to generate a query model that is good at discovering documents belonging to these clusters instead of the irrelevant one [12]. In Automatic Query Expansion (AQE), terms are given new weights to score the terms. Sum of weights will represent final score of terms, which is statistically good for item selection. Pure statistical weights may not functionally useful to represent query terms. Different functions have been proposed to assign high scores to the terms that best discriminate relevant from non-relevant documents [10]. A disadvantage of Query Expansion is the inherent inefficiency of reformulating a query [13]. The query is expanded using words or phrases with similar meaning to those in the query and the chances of matching words in relevant documents are therefore increased. This is the basic idea behind the use of a thesaurus in query formulation [15]. To improve the relatedness of the terms to documents, lexical resources Thesaurus, WordNet or Dictionaries usage promising little improvements in search results [16]. While global analysis mechanisms are inherently much more efficient than local ones (only dictionary lookups are performed during query time, rather than costly document retrieval and parsing), they are also likely to be less successful [1]. Document expansion by modifying Vector Space is to bring closer the query Vectors [14]. Good thesaurus for whole language is difficult to obtain. Synonyms are used to extract from thesaurus [18] for query expansion. Expansion terms are selected based on query association, where queries are stored with documents that are highly similar statistically. Falk Scholer and others [17] claimed that adding query associations to documents improves the accuracy of Web topic finding searches by up to 7%, and provides an excellent complement to existing supplement techniques for site finding. The studies are showing that, the query expansion improves the results of Information retrieval system. Statistical relatedness may not work properly and choose correct alternate terms to reformulate the query. Document expansion during indexing reduces search time. In this paper we studied the effect of Query with and without Expansion versus Document with and without Expansion using Synset. Proposed work is proven to increase recall and precision as well. In few

cases like, document expansion, precision is inversely affected the results.

a) Preprocessing of Telugu Text

Telugu is derived from Brahmi family [], one of the Dravidian languages. Telugu is morphologically rich language and word conflation is very high. The language scripts are complex to process, because they are combined syllables when compared to English. So it is difficult to preprocess using language models like stemming, n-gram etc. Romanization called WXNotation standards aim at providing a unique representation of Indian Languages in Roman alphabet [27]. Internally each script is represented UNICODE standard. The Unicode Standard, Version 6.2 assigned a hexadecimal code point for Telugu Scripts in the Range of 0C00-0C7F [28]. In this paper implementation is done by converting text from WX-to-UTF1 and UTF-to-WX before and after processing. This process slower the results, but efficiency in terms of recall and precision are not influenced. Carrying task directly in Unicode give faster results and possible, but processing text in Unicode level is difficult for programming. Our future work is planned to directly process in Unicode to improve the results speed. WX notations for Telugu language are given in Table 1.

Table 1 : WX-Notation for Telugu Scripts

అ [a]	ఆ[A]	ఇ[i]	ఈ[I]	ఉ[u]
ఊ[U]	ఋ[q]	ఎ[e]	ఏ[eV]	ఐ[E]
ఒ[o]	ఔ[oV]	అం[aM]	అః[aH]	క[ka]
ఖ[Ka]	గ[ga]	ఘ[G]	ఙ[fa]	చ[ca]
ఛ[Ca]	జ[ja]	ఝ[Ja]	ఞ[Fa]	ట[ta]
ఠ[Ta]	డ[da]	ఢ[Da]	ణ[Na]	త[wa]
థ[Wa]	ద[xa]	ధ[Xa]	న[na]	ప[pa]
ఫ[Pa]	బ[ba]	భ[Ba]	మ[ma]	య[ya]
ర[ra]	ల[la]	వ[va]	స[sa]	శ[Sa]
ష[Ra]	హ[ha]	ళ[lYa]	క్ష[kRa]	ఱ[rY]

III. Query Expasnion

Terms supplied by user may not be sufficient to express the concept and match documents. Terms may be out of bounds or in different vocabulary. Out of bounds problem can be solved by user feedback system. Vocabulary mismatch is common problem in Information retrieval. Vocabulary mismatch is one is of the principal causes of poor recall in Information Retrieval. Indexers and searchers invariably choose different subset of words to specify a given topic, causing retrieval techniques based on lexical matching to miss relevant documents [19]. Expansion of query at search time is called run time query expansion. Query Expansion is a process of reformulating the root query by adding an optimal set of terms that improves recall and precision. The motivation for query expansion is rate of failure in retrieving relevant documents by simple queries. Various Query Expansion methods are in regular practice to improve the retrieval performance. Local Analysis and Global analysis.

a) Local Analysis

Initial search results of given query are analyzed and used to expand the query called local analysis. The top ranked documents were taken to change weights of query terms and repeat the search [20][21]. User judge the relevance of top ranked items to the query as Relevance Feedback [8]. The thought of relevance feedback is to involve the user in the retrieval process so as to improve the final result set. The user issues an initial query. The system returns an initial set of relevant documents. In particular, the user gives feedback on the relevance of documents in an initial set of results. The system computes a better representation of the information need based on the user feedback [22]. It may cause the user to endure the process. Pseudo Relevance Feedback (PRF) is viable alternate to void user interaction during feedback. PRF is also called Blind Relevance Feedback or Automatic Relevance Feedback method, which automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an extended interaction. PRF via query-expansion has been proven to be effective in many information retrieval (IR) tasks [23]. In most existing works, the top-ranked documents from an initial search are assumed to be relevant and used for PRF. One problem with this approach is that one or more of the top retrieved documents may be nonrelevant, which can introduce noise into the feedback process. For all query expansion methods, pseudo relevance feedback (PRF) is attractive because it requires no user input [24]. Major problem with local analysis is that queries have an increased risk of query drift, as the top ranked documents are assumed to be relevant, while they may in fact not be [21]. In this paper we studied both Relevance Feedback and Pseudo relevance Feedback methods on a limited corpus. Top one document is considered as relevant and its terms are given more weight in-line with query terms and repeated search on same collection. Even though, sometimes original queries are totally modified with new terms and missing the concept of original query. Still it is found to be the best approach among alternate methods including global analysis, which is discussed in the next subsection.

b) Global Analysis

The global analysis considers term co-occurrences and their relationships in the corpus as a whole, which is used to expand the query independent from the old query. Expansion by global analysis does not rely on initial query terms and the results retrieved from it, so that refinements in the query will cause the new query to match other semantically similar terms. A common problem with these query expansion methods is that the relationships between the original query terms and the expanded query terms are not considered [25]. In this paper our directions are to use synset words for query expansion and study the effect on training corpus. Recall ratio is improved in this direction.

i. *Synset based Query Expansion*

Our research is continuing in Information Retrieval in Indian Languages, as Telugu is one of the most spoken languages in India as well as all over the world. Language resources are limited in Telugu language and cross language attempt are facing many challenges, where the features are different from one language to other language. For this work we collected and manually created Telugu-Telugu synset of <<<>>> for whole corpus consisting of around 3 laksh words. Query preprocessing is done using similar process as applied to document indexing in section 4.1. Query Terms are expanded using Synset and combined terms with OR Boolean operator. Expanded query is used for search on static Indexed documents. The results were given in Section 5. Somehow vocabulary miss match problem is addressed by synset inclusion through run time terms expansion. E.g. one word query is అమ్మ [amma] – mother. First the root word is verified with word corpus of dictionary look-up, if found all synonyms from synset are connected using OR operation to generate new query. i.e (అమ్మ [amma] OR మాత [mAwa] OR తల్లి [walli] – synonyms of mother in Telugu) treated a single term and weighted accordingly in query vector. This works good as we collected all possible synonyms in the corpus. Recall is greatly improved, at the same time precision is compromised due to deviating the concept as well as query drift.

ii. *Synset based Document Expansion*

It is impossible to predict the query from user and terms used by him. Instead of expanding terms during run time, as user need to wait for results, index terms of document set are expanded during indexing using synset. Off course the practice is not new, even though it is new to Indian languages especially for Telugu. Telugu Information Retrieval suffers from language resources. There is a demand for language resource to be developed for all Indian languages for public use. We created a synset for our training corpus of 40000 words with 1.375 synonyms in an average and this process is continuing to create for 1 lakhs words synset. All unique root words from entire corpus are extracted and created a dictionary file. A hash is maintained to list synonyms of a document term to match against query term during searching. Similar to Query Expansion this attempt deprecate the process and resulted precision loss.

c) Relatedness Measurements

Relevance Feedback: Vector Space Model

$$Qi = (q1, q2, q3, ...qn)$$

Q_i → is an initial query as a vector of terms q_j.

q_j → weight of each query term j in Q_i

A New query with added terms from the top retrieved documents D is given as expanded query to research. Weighting can be taken either Boolean value, in which 0 represents deletion of old terms and 1 represents addition of new terms to the query.

$$Di = (d1, d2, d3...dn)$$

D_i → is an top Document as a vector of terms d_j.

d_j → weight of each Document term j in D_i

Similarity of Query and document is measured by:

$$Sim(Qi, Di) = \sum_{i=0}^{n} qi \bullet di$$

Rocchio [8] proposed a Relevance Feedback algorithm which better suggest a new query as:

$$Q_{new} = \alpha Q_{old} + \beta \sum \frac{Dr}{|Dr|} - \delta \sum \frac{Dnr}{|Dnr|}$$

Where Qnew is Reformulated query and Qold is initial query with Dr Relevant returned Documents, |Dr| number of relevant documents. Dnr is non-relevant returned documents and |Dnr| is total non-relevant documents in terms of vectors. With α is original query weight, β is related document weight and γ is weight of non-relevant documents. Less importance terms are represented with 0 in Boolean vector models. Concept of a query may depends on less weighted terms too, hence it is important to equally consider less weighted terms in sorting order. An alternate term weighting method call probabilistic approach better serve the purpose. The documents are ranked in decreasing order of rank as per the expression:

IV. Implementation

Telugu language resources are limited for research. We collected 3500 Documents from daily news portals and manually categorized into 10 categories as shown in Table 1. Initially all documents are kept under one set and run the search using 10 queries and followed by search against categorical sets of documents. There is no difference in results as

documents are properly indexed before running search. Little search time varies from search on whole collection and categorical collection of documents. If the documents were categorized the results were bit faster. This time factor is important, but our aim is to improve the precision and recall.

Table 2 : Categorical Documents collection for testing

#Queries 10			
S.No	**Category**	**#Docs**	
1	Business	150	**Total #Docs 3500**
2	Devotional	1552	
3	Editorial	150	
4	Historical Places	152	
5	Literature	305	
6	Politics	332	
7	Science	152	
8	Songs	298	
9	Sports	294	
10	Stories	155	

a) Indexing and Searching using Synset

1. Collect set of Documents manually or Call Web Crawler to collect.
2. Clean the Document and Store in Text format.
3. Tokenize the Item and extract all words.
4. Eliminate the Stop words based on POS Tagging or by comparing collections of corpus stop list and maintain list of unique terms.
5. Apply linguistic process to get root words as index terms (We applied Morphological Analyzer to extract language features like POS Tags & Roots).
6. Give weight to each term in the document and query vector to represent the importance of the term for expressing the meaning of the document and query. There are two widely used factors in calculating term weights.
   i. Term frequency (tf): Occurrence of a term i in document j calculated by:

$$tf_{i,j} = \frac{fq_{i,j}}{\max(fq_{i,j})}$$

Where fqi,j frequency of term i in document j.

   ii. Create Inverted List [26] consisting of Document Ids and term frequency against Dictionary lookup. Inverse Document Frequency a term i is calculated using idf_i

$$idf_i = \log \frac{N}{ni}$$

Where N total no. of documents and ni is number of document that term i occurs.

7. Synset is used to identify synonyms of a term in a document, if found term frequency is incremented in Inverted List to give more importance to that file containing a term with synonyms. Weighting factor are greatly affected by synset.
8. Relevance judgment is find using cosine similarity measure between document and query vectors:

$$\cos\theta = similarity(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \bullet |\vec{q}|}$$

$$\cos\theta = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

Where Θ is angle between *i j i q w w* , , & , wi,j is weight of term i in document j and wi,q is weight of term i query q. Θ varies from 0 to 1.

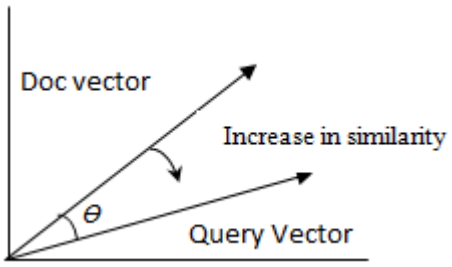


Figure 1 : Cosine Similarity

9. Use of Hash Table to represent Synset gives faster access.

b) Boolean Retrieval

In Boolean model proposed by Baeza Yates [4] represents entire document as bag of indexed words. Set of terms in a Boolean query are connected by AND, OR, NOT operators. When huge collection of documents are matched, Ranking of more relevant documents reduce the overhead on user to locate expected out come out of top ranked documents. This method works only for exact matches e.g. www.academic.research.microsoft.com which is treated as expert search engine, but general search engines like google, yahoo, bing etc gives related search too. So Vector space model is taken in this paper to search with one Boolean operator OR to connect multiple synonyms in reformulated query at search time.

IV. Results Analysis

In this paper we investigated affect on precision and recall when query is connected with synset. Figure 2 is a simple precision – recall graph as baseline approach to compare the proposed system. Fig.2 and Fig. 3. Shows precision-recall in normal search process without any aids. Query expansion is applied by

analyzing top one document as relevant and adding new terms to the query. This feedback is iterated once and results are plotted in Figure 2.

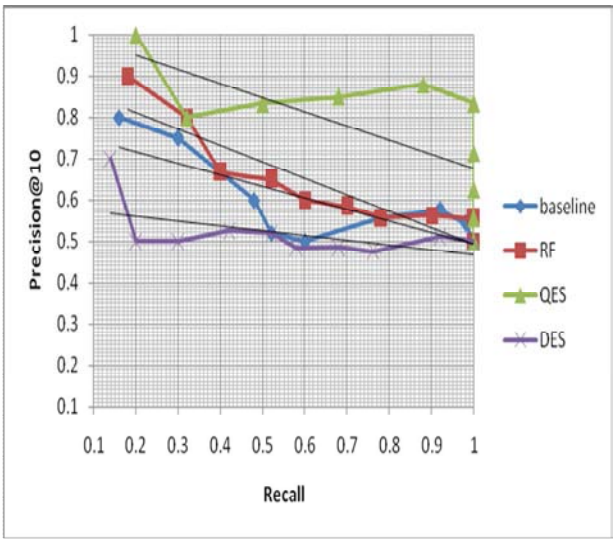


Figure 2 : Recall – Precision@10 variations with baseline, Relevance Feedback (RF) by top 1 document, Query Expansion using Synset (QES) and Document Expansion using Synset (DES)

Recall is not affected much in Figure 3., but precision loss is observed. All methodsare comapred in Figure 4 with Precesion@10. Baseline metho Query Expansion using synset is plotted in figure 4. When we use synset for query expansion instead of relevance feedback method, the precision is improved along with recall. Whereas Document Expansion with synset instead of Query Expansion, the results were greatly affected by both precision and recall. Query expansion with synset outperforms among all methods and we argue that, use of synset to expansd query is better than document expansion. The proposed system has to be tested on huge corpus so as to claim in universal Information Retrieval. Experiments are taken on TREC using similar methods by meny researchers and found precision loss. As there is no standard corpus for Telugu language we tested on private corpus developed by us.

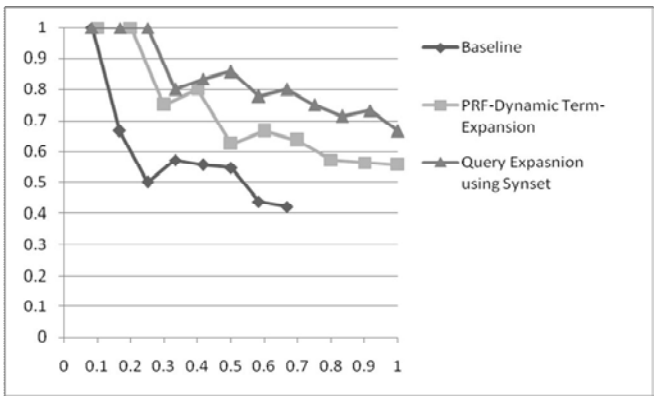


Figure 3 : Term Expansion using Semantic Network called Synset at Different levels

Table 3 : Growth of Precision(P) and Recall (R)

Method	%P @ 10	% P- Growth	% R	%R- Growth
Baseline	**60.10**	NA	**61.4**	NA
RF	**63.80**	3.7	**64.00**	2.6
QES	**75.90**	15.8	**75.80**	14.4
DES	**52.00**	-8.1	**55.20**	-6.2

Terms Expansion during indexing using synset gave poor performance as shown in Table 3. The search results are bit faster in DES when compared to RF & QES methods, but these to are good in relevance calculation.

V. Conclusion

Indian is multilingual country stands in 2nd in population. There is observable growth in literacy, but people prefer to use local languages after English. There is necessity for cross lingual information retrieval systems to serve the users according to their information needs. Most of the Indian Languages are having unique language features and it is difficult to translate from one language to other, even Google search engine fails to produce exact cross lingual results. Building monolingual information retrieval is a mandatory task, where compatibility may not be an issue in using language resources. Once identifying all features of a language, it is easy to translate into other language by mapping rules. Information Retrieval in Telugu Language is in inception level, due to lack of language resources like POS Taggers, Entity Recognizers, Morphological Analyzer, Dictionaries, WordNet, Ontologies et. al. The results in this paper were given hope to continue with query expansion. Use of controlled vocabulary may further improve the results. Anyhow Query expansions techniques will have inverse effect on precision and improvement in recall. For the end user precision is more important, as he expects results to be displayed on top in once shot. Exploring concept of the query using synset or WrodNet may give better performance. We need to investigate how Information retrieval system in Telugu Language works by using query concepts.

References Références Referencias

1. Bodo Billerbeck Justin Zobel, 2005. Document Expansion versus Query Expansion for Ad-hoc Retrieval 10th Australasian Document Computing Symposium, Sydney.
2. Ramakrishna Kolikipogu and Padmaja Rani B, 2012. Reformulation of Web query using Semantic Relationships, International ACM-ICACCI-12. [3] Y.Y. Yao, 2002. Information Retrieval Support.

3. Systems, FUZZ-IEEE'02 -- The 2002 IEEE World Congress on Computational Intelligence, Honolulu, Hawaii, USA, May 12-17, 2002, 1092-1097.
4. Baeza-Yates, R. and Ribeiro-Neto, B, 1999. Modern Information Retrieval, Addison Wesley, New York.
5. Dominich, S. 2001. Mathematical Foundations of Information Retrieval, Kluwer Academic Publishers, Dordrecht.
6. Dan I. Moldovan and Rada Mihalcea, 1998. Improving the Search on the Internet by using Word Net and Lexical operators. IEEE Internet Computing.
7. Maron, M. E. and Kuhns, J. L. 1960. On relevance, probabilistic indexing and information retrieval. J. ACM 7, 3, 216–244.
8. Rocchio, J. J. 1971. Relevance feedback in information retrieval. In The SMART Retrieval System, G. Salton Ed., Prentice-Hall, Englewood Cliffs, NJ, 313–323.
9. Ruthven I, 2003. Re-examining the potential effectiveness of interactive query expansion. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in information Retrieval. ACM Press, 213–220.
10. Claudio Carpineto and Gioovanni Romano, 2012. Automatic Query Expansion in Information Retrieval, ACM Computing Surveys, Vol. 44, No. 1, Article 1, Publication date: January.
11. C. Zhai and J. Lafferty, 2001. Model-based feedback in the language modeling approach to information retrieval. In Proceedings of the 10th international conference on information and knowledge management, pages 403–410.
12. Bin Tan, Atulya Velivelli, Hui Fang, ChengXiang Zhai, 2007 Term Feedback for Information Retrieval with Language Models, SIGIR-2007 Proceedings, 263- 270.
13. Bodo Billerbeck and Justin Jobel, 2005. Document Expansion versus Query Expansion for Ad-hoc Retrieval, Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, 34-41.
14. G. Salton, 1971. The SMART Retrieval System. Experiments in Automatic Document Processing, Prentice-Hall, Englewood Cliffs, 373.393. NJ.
15. J. Xu and W. B. Croft. 2000. Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems, Volume 18, Number 1, pages 79.112.
16. Jinxi Xu and W.Bruce Croft, 1996. Query Expansion using Local and Global Document Expansion, ACM SIGIR Conference Proceedings, 4-11.
17. F. Scholer, H. E. Williams and A. Turpin, 2004. Query association surrogates for web search. Journal of the American Society for Information Science and Technology, Volume 55, Number 7, 637-650.
18. K. Sparck Jones, 1971. Automatic Keyword Classification for Information Retrieval. Butterworths, London.
19. George W Furnas, 1988. Information retrieval using Singular Value Decomposition Model of Latent Semantic Structure, ACM-SIGIR-1988, 465-480.
20. Attar, R., & Fraenkel, A. S., 1977. Local Feedback in FW-Text Retrieval Systems. Journal of the Association for Computing Machinery, 24(3), 397–417.
21. Croft, W. B., & Harper, D. J., 1979. Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35, 285-295.
22. Kolikipogu Ramakrishna, Dr. B. Padmaja Rani, 2010. Information Retrieval in Indian Languages: Query Expansion models for Telugu language as a case study, 4th IEEE - International Conference on Intelligent Information Technology Applications, china.
23. Yang Xu, Gareth J. F. Jones, 2009. Query Dependent Pseudo Relevance Feedback based on Wikipedia, Association of computer Machinery SIGIR'09 Conference Proceedings.
24. C Buckley, G. Salton, J. Allan, and A. Singhal. 1994. Automatic query expansion using SMART: TREC-3.
25. Renxu Sun Chai-Huat Ong Tat-Seng Chua, 2006. Mining Dependency Relations for Query Expansion in Passage Retrieval, SIGIR'06, Seattle, Washington, USA, 382-389.
26. C. D Manning, P. Raghavan, and H. Sch utze, 2009. An introduction to information retrieval. Cambridge University Press.
27. Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 2010. Natural Language Processing: A Paninian Perspective. PHI,
28. Ramakrishna Kolikipogu and Padmaja Rani B, 2013. Study of Indexing Techniques to Improve the Performance of Information Retrieval in Telugu Language, IJETAE, Volume 3, Issue 1.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Ontology Mapping for Cross Domain Knowledge Transfer

By Santosh Kumar Banbhrani, Xu DeZhi & Mir Sajjad Hussain Talpur
Central South University

Abstract - The proliferation of domain specific ontologies has improved the ability to represent process and store information in regard to highly specialized domains. However, adhoc transfer of information between domain specific ontologies is not possible. Consequently, multiple solutions have been pro-posed and evaluated as means of facilitating the adhoc transfer of information between another. These range from, structural approaches, which attempt to match knowledge structures between ontologies; lexicographical approaches, that use high level reasoning to match concepts between related ontologies and finally, local structure approaches which look for similar local structures between ontologies to facilitate the transfer of information. To date, the success rate of the published algorithms has been relatively poor. Some of the most successful algorithms, at best are able to match around 50% of the concepts between related ontologies. In this paper we propose a novel global-local hybrid approach to improve the success and accuracy of adhoc information transfer between domain specific ontologies. We demonstrate the efficiency of the proposed algorithm by matching the nodes of three inter-related medical domain ontologies. This demonstrates a significant improvement over existing lexicographical and structural approaches.

Keywords : *domain knowledge, heterogeneity, ontology mapping, semantic web.*

GJCST-C Classification : *D.2.11, D.2.13*

Strictly as per the compliance and regulations of:

Ontology Mapping for Cross Domain Knowledge Transfer

Santosh Kumar Banbhrani [α], Xu DeZhi [σ] & Mir Sajjad Hussain Talpur [ρ]

Abstract - The proliferation of domain specific ontologies has improved the ability to represent process and store information in regard to highly specialized domains. However, adhoc transfer of information between domain specific ontologies is not possible. Consequently, multiple solutions have been proposed and evaluated as means of facilitating the adhoc transfer of information between another. These range from, structural approaches, which attempt to match knowledge structures between ontologies; lexicographical approaches, that use high level reasoning to match concepts between related ontologies and finally, local structure approaches which look for similar local structures between ontologies to facilitate the transfer of information. To date, the success rate of the published algorithms has been relatively poor. Some of the most successful algorithms, at best are able to match around 50% of the concepts between related ontologies. In this paper we propose a novel global-local hybrid approach to improve the success and accuracy of adhoc information transfer between domain specific ontologies. We demonstrate the efficiency of the proposed algorithm by matching the nodes of three inter-related medical domain ontologies. This demonstrates a significant improvement over existing lexicographical and structural approaches.

Keywords : domain knowledge, heterogeneity, ontology mapping, semantic web.

I. Introduction

Ontology's have become a valuable tool to help quantify and process information for decision support systems in highly specialist knowledge domains. Consequently large amounts of both qualitative and quantitative data are processed and stored [3] in various expert systems. The drawback of using these ontologies is that, automated transfer between the systems requires extensive operator intervention in the form of specialist data transfer tools. These tools require the designer to manually map the common information concepts between the two ontologies. As the complexity of the data stored an ontology increases, the complexity of the mapping task and the probability of an error increases. One recent study published by Oellrich et al.[2] found that formal mapping between two ontology's representing the same knowledge domain (Human Pheno-type Ontology and Mammalian Phenotype Ontology defined using the phenomeblast software) was successful at mapping only 48% of the concepts between ontology's. This lack of success at mapping between ontology's has multiple underlying factors such as, knowledge conceptualizations by the developers with implicit assumptions and/or conflicting knowledge structures due on developer assumptions. The assumptions underlying the development of ontology definition and structure arise out of a lack of external standards for the knowledge domain being modeled. External standard setting bodies represent a specific expression of the nature of the information being classified and are able to establish formal relationships for information stored in an ontology can only mitigate this challenge. Thus at the instance of definition, an ontology can at best represent a subset of the scientific world-view in regard to that knowledge domain. This problem is further exacerbated by the presence of multiple standard setting bodies. For example when developing an ontology for medical diagnosis support systems, the developers have a choice of at-least five medical terminology thesauri when using the English language. Individually these controlled vocabularies have well defined application areas with little or no overlap. However, when used to develop an ontology for a specific purpose (clinical diagnosis) the underlying assumptions and world-views of the thesaurus chosen guide and inform the structure of the ontology. This acts as an impediment to the transfer of information between ontologies based on different thesauri. Additionally, when developing an ontology for a specific application area, by choice, only a small subset of concepts in a domain will be used to create the ontology. Due to this, translating all the concepts between from one ontology to another is extremely unlikely to succeed. Therefore, success in concept translation will rely on being able to map all relevant concepts.

This document will report the results of a lexicographical and structural hybrid approach that has been found effective at mapping relevant concepts between related ontologies, developed using a well-defined and restricted vocabulary. This document is organized as follows, the next section will review existing literature for inter-ontology data transfer, following this the next section will present the results for when mapping between three medical domain algorithms with

Author α : Masters of Engineering in Computer Application Technology. E-mail : santosh.banbhrani@gmail.com
Author σ : Professor School of Information Science and Engineering. E-mail : hunan.xu@mail.csu.edu.cn
Author ρ : Ph.D. (Computer Science) Research Scholar. E-mail : mirsajjadhussain@gmail.com
Address : School of Information Science and Engineering, Central South University, Changsha Hunan, China.

techniques identified in literature. After this the next section will describe the novel algorithm proposed in this paper will be described. Finally the results after re-mapping the same three ontologies are presented, after this the conclusion identifies further work that is needed to validate this technique.

II. Related Work

As noted in the previous section, mapping between ontologies developed for limited vocabularies is an extremely active research area. Multiple techniques have been proposed and demonstrated as being effective at mapping between related ontologies; one comprehensive survey of ontology mapping tools published in 2006 by Choi, Song and Han [1] proposed that the terms "*ontology mapping*", "*ontology alignment*" and "*ontology merging*" refer to and indicate different approaches to solving a common subset of challenges.

The paper segments ontology mapping into the following subsets:

A global ontology and local ontologies Here the mapping between ontologies is used to query information from other ontologies, or to map a concept from one ontology into a view.

Mapping between local ontologies This is used to transform entities in one or more source ontology into entities in the target ontology.

Ontology merge and alignment Used to identify unique concepts found in one or more source ontologies being considered for merging or to identify redundant or overlapping concepts.

From the tools described in the paper, semantic matching was common to all the tools described in addition to semantic matching the following approaches were been used for mapping entities and concepts between ontologies. These include but are not limited to, hierarchical mapping, [4-5] probability distribution mapping [6].

Table 1 : Entity mapping success rate between the ontologoies

	Ontology 1	Ontology 2	Ontology 3
Ontology 1	**34**	**12**	**10**
Ontology 2	**12**	**40**	**13**
Ontology 3	**10**	**13**	**53**

Lexical mapping [7] and probabilistic pair matching [8-9]. The tools evaluated in the survey were mainly semi-automated and were designed to be used as support tools for human decision making when mapping entities and concepts between ontologies. Only one of the surveyed tools "CTX Match" [2] is a complexly automated algorithm. As the survey is now 10 years old, the need for a completely automated ontology mapping algorithm has become imperative.

The need for an automated tool has primarily grown due to two reasons, the increasing size and complexity of ontologies used in expert and decision support systems. Secondly, the need to migrate large amounts of accumulated data from obsolete systems to updated ontologies. Invariably for obsolete systems, underlying documentation may be missing, incomplete or unavailable due to various factors.

This work proposes a completely automated ontology mapping algorithm, therefore from the tools evaluated in the survey, the CTX Match algorithm proposed by Bouquet, Serafini and Zanobini8 will be directly comparable.

CTXMATCH is a hierarchical logical reasoning tool that uses the hierarchical relationship between the entities in both the target and source ontologies. The mapping between the source and target inputs H, and H1 in HCs, and for each pair of concepts
(a node with relevant knowledge including meaning in Hierarchical classifications), returns their semantic relation . For example, k is more general than , k is less general than , k is equivalent to , k is compatible with
and k is incompatible with .

a) CTXMATCH Results

After processing the three ontologies with the CTXMATCH algorithm the following results were obtained.

Subsequent research into algorithmic ontology mapping has improved on hierarchical mapping by using techniques derived from directed graph matching. To illustrate the potential of a generic directed graph node matching technique, the same three ontologies were re-mapped. Comparing the results in 1 and 2, we can see that even without semantic matching the directed graph entity matching technique is more effective when mapping between ontology 1 and ontology 3. The next section consist of the following the problem statement, introduction to directed graph matching and finally the algorithm description.

Table 2 : Directed graph entity matching

	Ontology 1	Ontology 2	Ontology 3
Ontology 1	**34**	**07**	**17**
Ontology 2	**07**	**40**	**10**
Ontology 3	**17**	**10**	**53**

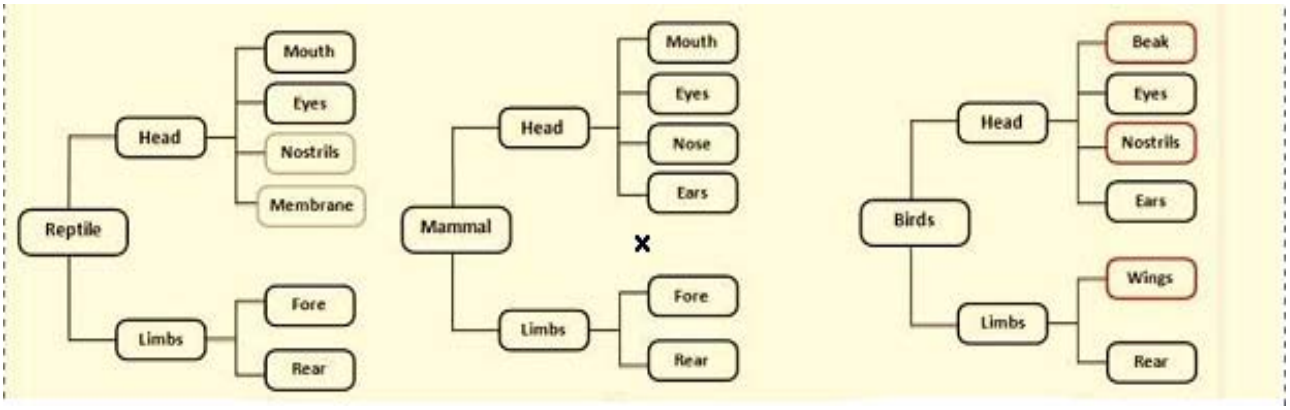
III. Problem Statement

The work in this document is based from a study comparing the performance of sophisticated algorithms evaluated as part of The Ontology Alignment Evaluation Initiative to a simple lexicographical ontology mapping algorithm. From these documents the following common themes can be identified:

- Mapping between domain specific ontologies is a challenging problem, for which currently manual concept mapping is the only effective solution.
- Pattern matching and machine learning algorithms are reasonably successful at ontology mapping. As the assumptions and world views that are a factor into ontology development are di cult to quantify, there is likely to be an upper limit to the concept mapping accuracy.
- Mapping between ontologies using limited vocabularies for similar use-cases is more likely to be effective and accurate.

a) Directed Graph Matching Example

As identified in existing literature lexicographical ontology mapping techniques are extremely effective especially for limited vocabularies. Using a zoological reference textbook as a sample vocabulary for the ontologies in figure 1. A reasoner success-fully maps the *beak* class to the mouth class. The vocabulary used will introduce uncertainty in the mapping of the classes *nostrils* and *membrane* classes of the rep-tile ontology and the *wings* class of the bird ontology. Therefore, the number of mapped classes will range from 18 to 22, i.e. 81% to 100%.



(a) (b) (c)

Figure 1 : Sample Ontologies

On the other hand for example, using a directed graph technique to map in-formation between the classes is extremely simple as the ontologies have the same structure.

Table 3 : Simulated lexical ontology entity matching for figure 1

	Reptile	Mammal	Birds
Reptile	**09**	**07**	**08**
Mammal	**07**	**09**	**06**
Birds	**08**	**06**	**09**

Table 4 : Simulated directed graph matching for figure 1

	Reptile	Mammal	Birds
Reptile	**09**	**09**	**09**
Mammal	**09**	**09**	**09**
Birds	**09**	**09**	**09**

Therefore this map will be 100% successful at class mapping. If on the other hand the structure of the ontologies were to be modified to reflect a different world view, as illustrated in figure 2. For this sample, the graph based method would fail when mapping the world view, as illustrated in figure 2. For this sample, the graph based method would fail when mapping the "*head*"information between the bird and the other two ontologies. Therefore of the 22 classes only 19 classes are successfully mapped, 86% success rate.

These results are summarized in tables 3, 4 and 5.

IV. Proposed Work

To reduce the uncertainty in the lexicographical approach, a novel combined ontology mapping algorithm was proposed. The algorithm combines the lexicographical mapping with the directed graph approach to reduce mapping uncertainty.

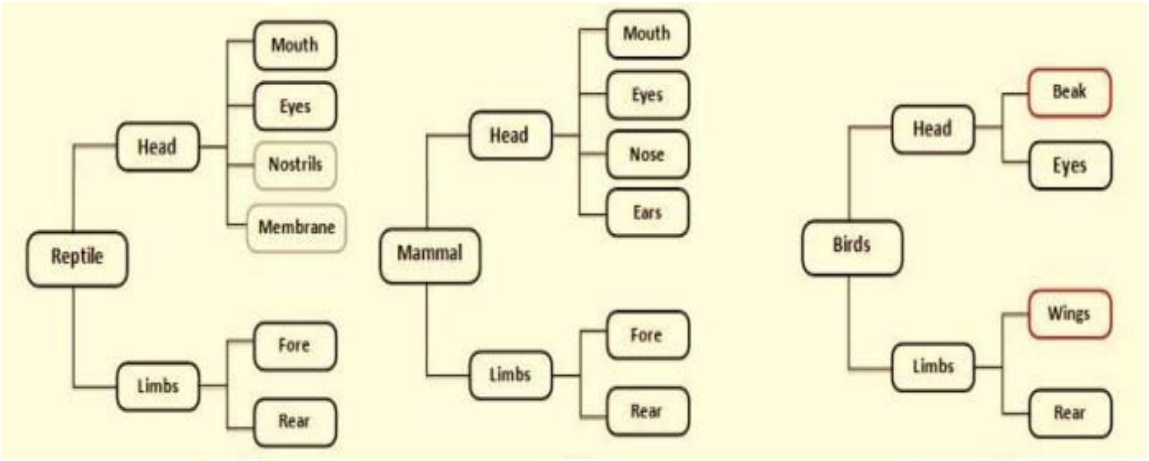
1. Use a thesaurus based synonym (lexicographical) search to identify concept commonality and term networks in the two domains.
2. Read class structure for source and target ontologies to generate node-edge graphs to identify common class structures.
3. Use value matching to bootstrap and validate structural mapping.
4. Use word networks to map areas that do not match structurally and re-evaluate parent nodes.
5. Repeat steps three and four to find any updated root nodes that:

Table 5 : Simulated lexical ontology entity matching for figure 2

	Reptile	Mammal	Birds
Reptile	**09**	**09**	**07**
Mammal	**09**	**09**	**07**
Birds	**07**	**07**	**09**

Table 6 : Results of processing with proposed algorithm

	Ontology 1	Ontology 2	Ontology 3
Ontology 1	**34**	**24**	**23**
Ontology 2	**24**	**40**	**32**
Ontology 3	**23**	**32**	**53**



(a) (b) (c)

Figure 2 : Sample Ontologies

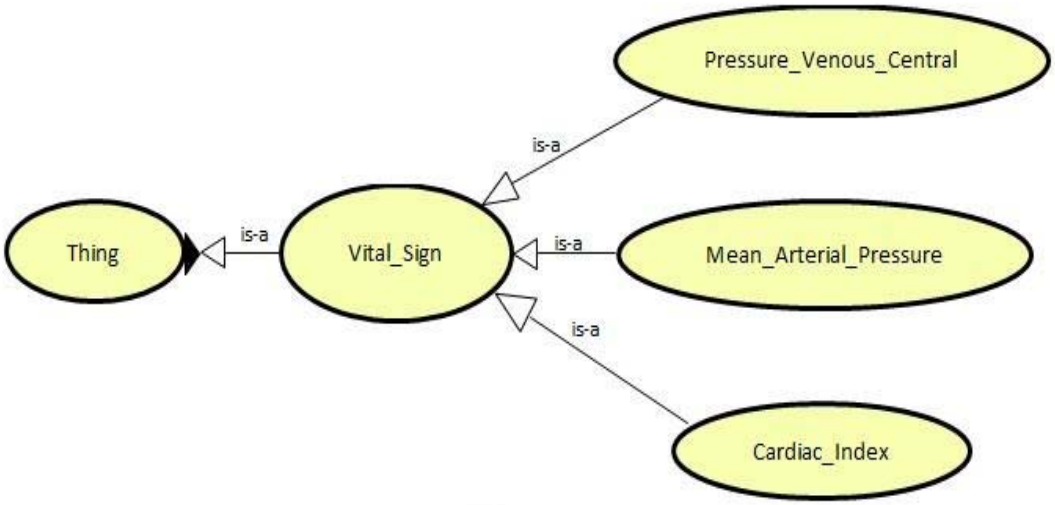
The limited vocabulary dictionaries that were used for the lexicographical matching were obtained from the nih.gov website. To generate the lexicographical map, the Apache-NLP libraries [10] in java were used to generate word associations between the three dictionaries. To generate the `terms of interest' networks of descriptor terms are generated for words common to all three directories.

Using the direct graph matching technique illustrated in the previous section (section??) is used as follows. For a root node pair $(R—N)$ in ontology 1 and a similar root-node pair $(R'—N')$ in ontology 2. Node pairs that have the same structure i.e. same properties such as scientific units (physical, chemical or biological), data types are considered matches. For root's, with more than one nodes, an arbitrary value (experimentally determined to be .75) is used as a threshold. That is, if more than 75% of the child nodes of a node match the child nodes of a root node of the target ontology then the root nodes are considered a match.

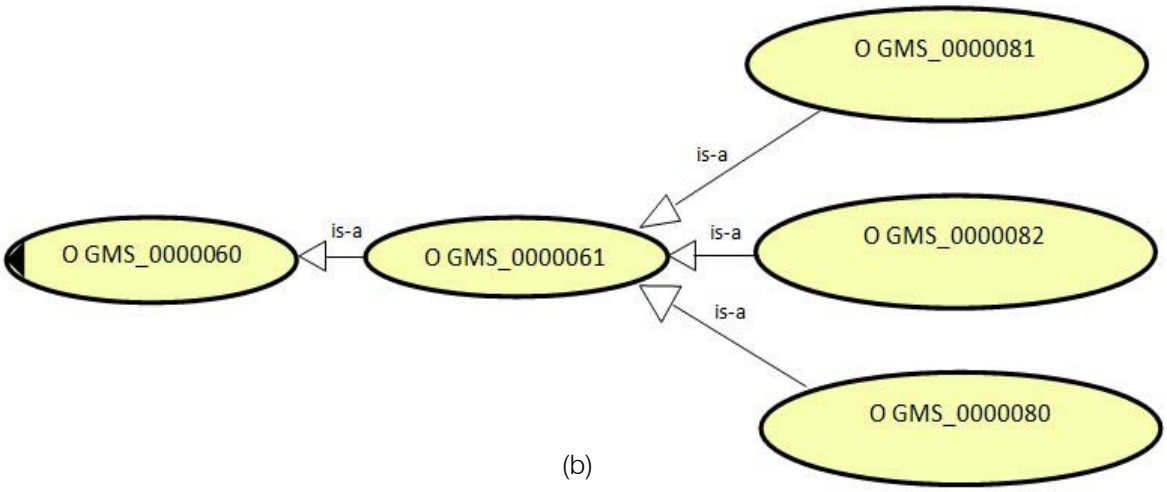
Following this, any nodes in the source and target ontologies that do not match. NLP network search is used to find matches any nodes that initially were not found to have any corresponding matches. Any subsequently matching nodes are marked as such and the nodes are revaluated to identify any root nodes that may now meet the threshold for matching child nodes. The results of processing the three ontologies and with the proposed algorithm are detailed next.

V. Results

As we can see in the table 6, the proposed algorithm improves significantly when compared to the hierarchical or directed graph matching as illustrated in tables 1 and 2. One reason for this could be because all three ontologies are medical support ontologies that use a significantly constrained vocabulary. This and the availability of comprehensive dictionaries that the JAVA NLP toolkit has been designed to process, probably make these ontologies non exemplars when identifying drawbacks to this approach.



(a)



(b)

Figure 3 : Successful Combined Match

VI. Conclusion

Since many ontology mapping algorithms have been proposed, a group of criteria are urgently needed to evaluate and compare the results of different algorithms. However, former measures have their own limitations, and none of them can guarantee that semantically equivalent alignments always score the same, which should be a basic character of a real semantic evaluation. By this we have demonstrated that the proposed algorithm can be very effective then existing algorithms. Their performance is equivalent to the performance of the more innovative algorithms. Our evaluation has validated that most of the progressive algorithms are either not freely available or do not scale to the size of biomedical ontologies. We have tested this algorithm and got result 2 which is better than the result 1 which is based on existing algorithms which I used as part of our algorithm. Now that we have some preliminary significant results demonstrating the effectiveness of this approach for use with medical support ontologies. The effectiveness of this algorithm needs to be evaluated with larger and more complex ontologies. In future work we will focus upon testing with ontologies of greater size. Those tests will provide for solider proof whether this method can be successfully applied to the ontology integration problem.

Acknowledgment

This work is supported by National Natural Science Foundation of China under grant No.60970096 and 90818004, the Postdoctoral Science Foundation of China under Grant No.20080440988 and the Natural Science Foundation of Hunan Province, China Under Grant No.09JJ4030.

References Références Referencias

1. Namyoun Choi, Il-Yeol Song, and Hyoil Han. A survey on ontology mapping. ACM Sigmod Record, 35(3):34 {41, 2006}.
2. Anika Oellrich, Georgios Gkoutos, Robert Hoehndorf, and Dietrich Rebholz-Schuhmann. Quantitative comparison of mapping methods between Human and Mammalian Phenotype Ontology. Journal of Biomedical Semantics, 3 (Suppl 2):S1, 2012.
3. L. Yao, A. Divoli, I. Mayzus, J.A. Evans, and A. Rzhetsky. Benchmarking ontologies: bigger or better? PLoS Computational Biology, 7(1): e1001055, 2011.
4. Doan, A., Domingos, P., & Halevy, A. (2003). Learning to match the schemas of data sources: A multistrategy approach. Machine Learning, 50(3), 279-301.
5. Paolo Bouquet, Luciano Sera ni, Stefano Zanobini, Semantic Coordination: A New Approach and an Application", ISWC 2003, LNCS 2870, pp. 130-145, 2003.
6. AnHai Doan, Jayant Madhavan, Pedro Domingos, Alon Halevy\Learning to Map between Ontologies on the Semantic Web", VLDB Journal, Special Issue on the Semantic Web, 2003.
7. John Li, \LOM: A Lexicon-based Ontology Mapping Tool", Proceedings of the Performance Metrics for Intelligent Systems (PerMIS. '04), 2004.
8. Mitra, P and Wiederhold, G, \Resolving Terminological Heterogeneity in Ontologies", Proceedings of the ECAI'02 workshop on Ontologies and Semantic Interoperability, 2002.
9. Prasenjit Mitra, Natasha F. Noy, Anju Jaiswals\OMEN: A Probabilistic Ontology Mapping Tool" International Semantic Web Conference 2005: 537-547
10. opennlp.apache.org

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

An Intelligent Method of Secure Text Data Transmission through Internet and its Comparison using Complexity of Various Indian Languages in Relation to Data Security

By Devasish Pal, Padiga Raghavendra & Dr. A Vinaya Babu
Sri Sai Jyoth Engineering College

Abstract - Security of data transmitted through internet has posed a number of challenges. Data transmitted can be in the form of text, pictures, audio and video clips. In this paper a study has been carried out to find the relationship between the complexities of various Indian languages and its relation to text data security through an intelligent method of converting the text data before transmission. Complexity has been determined from the percentage retrieval by cryptanalyst using reverse frequency mapping without knowing the key. Percentage retrieval with and without converting the text data into an intelligent intermediate form have been compared and repeated in many languages along with English. The percentage retrieval of encrypted data using a language other than English, after converting the same data into a coded file and a dictionary is almost negligible.

Keywords : *complexity, dictionary, compression, frequency, retrieval, occurrence, coded file.*

GJCST-C Classification : *E.3*

Strictly as per the compliance and regulations of:

An Intelligent Method of Secure Text Data Transmission through Internet and its Comparison using Complexity of Various Indian Languages in Relation to Data Security

Devasish Pal [α] , Padiga Raghavendra [σ] & Dr. A Vinaya Babu [ρ]

Abstract - Security of data transmitted through internet has posed a number of challenges. Data transmitted can be in the form of text, pictures, audio and video clips. In this paper a study has been carried out to find the relationship between the complexities of various Indian languages and its relation to text data security through an intelligent method of converting the text data before transmission. Complexity has been determined from the percentage retrieval by cryptanalyst using reverse frequency mapping without knowing the key. Percentage retrieval with and without converting the text data into an intelligent intermediate form have been compared and repeated in many languages along with English. The percentage retrieval of encrypted data using a language other than English, after converting the same data into a coded file and a dictionary is almost negligible.

Keywords : complexity, dictionary, compression, frequency, retrieval, occurrence, coded file.

I. Introduction

Exponential growth of the internet and free accessibility to all users across the globe, security of data across internet has become a prime concern. This security aspect can be further divided into security of data and information in individual systems and in transit between internet users across the network. Data and information can be further divided into text and non text data eg pictures, graphics, audio and video clips. Earlier the text data used to be only in English language. Introduction of unicode and the process of localization [2] encouraged the information exchange of language based context resulting in text data being transmitted in all languages across the internet. This paper deals with security of text data while being transported across the network. To achieve security of data transmission, cryptography is one of the methods in which the security goals can be achieved by means of encryption and decryption. The key used for cryptography can be symmetric or asymmetric key. The encryption can be of blocks of fixed/variable size bit stream transformed to cipher stream. They use either block cipher or stream cipher techniques for transformation. Parameters in these schemes are mainly algorithm and key. Larger the key size, greater is the security of data and slower is the data rate. One more parameter has been considered i.e the complexity of a language [6] with a case study on Telugu. Greater the complexity of the language, greater is the security of text transmitted in that language keeping other parameters like encryption algorithms and the key constant. A simple logical conclusion is that if the text of a script is complex then the same level of security can be achieved with lesser key size. Subsequently a comparative study has been carried out over English and Telugu with Bengali as a case study [7] and it was observed that percentage retrieval of data in Bengali is less than Telugu and English.

In this paper, a comparative study has been carried out on various other Indian languages and adding a fourth security parameter i.e an intelligent method of text data encryption with security [8].

II. Review

A lot of study has gone into making the job of cryptanalysts simpler. Different languages in the world consist of characters displaying different properties and behavior [3, 4] which help in the process of cryptanalysis. One of the methods of determining the language complexity is by the frequency analysis. In this process frequency of each symbol in the encrypted message is determined. This information is used by cryptanalysts, to determine which cipher text symbol maps to the respective plaintext symbol. In transposition systems, the letter frequencies of a cryptogram are identical to that of the plaintext. In the simplest substitution systems, each plaintext letter has one cipher text equivalent. The cipher text letter frequencies are not identical to the plaintext frequencies, but the same numbers will be present in the frequency count as a whole. A method for fast cryptanalysis of substitution ciphers has been proposed by Thomas Jakobsen [1] which uses the knowledge of diagram distribution of the cipher text. The individual letters of any language occur

Author α : Associate Professor, IT SSJEC, Hyderabad.
E-mail : dpal55@gmail.com
Author σ : Student, III year IT SSJEC, Hyderabad.
E-mail : raghavendra.padiga@gmail.com
Author ρ : Principal JNTUH. E-mail : dravinaybabu@yahoo.com

with greatly varying frequencies [9]. This factor has been used to solve varying simple ciphers. There are two general approaches to solve simple ciphers. One makes use of the frequency characteristics and the other uses the orderly progression of the alphabet to generate all possible decipherments from which the correct plaintext can be picked up. Statistical analysis of the frequencies of multiple letters when compared to single letters have been found to be more helpful while retrieving part of plain text message. By using the combined techniques of monogram frequencies, keyword rules and dictionary checking the cryptanalytic technique of enhanced frequency analysis has been developed [5].

Plain text is encrypted using the proposed algorithm resulting in cipher text. The frequencies of different characters in the cipher text are extracted. Mapping is carried out between the characters of plain text and cipher text based on these frequencies. Now the characters in cipher text are replaced with the mapped characters of plain text and the percentage of the exact retrieval as compared to plain text is calculated by K.W. Lee et.al [5].

III. Conditional Probability

A vast study has been carried out into the frequency of occurrence of characters of many languages. The characters of different languages have different frequency patterns. This information helps a cryptanalyst to retrieve data from a cipher text by reverse frequency mapping. The percentage of data retrieved increases with the increase of corpus size of a sample text. As an example let us take a case study of English language. First a corpus frequency string is calculated with a very large corpus of English text. Corpus frequency string consists of all the different characters available in the corpus text. Next the percentage of occurrence of each character of corpus frequency string is calculated. To find the percentage retrieval of text after encryption from a new sample text, the new sample text is encrypted. The cryptanalyst using reverse frequency mapping tries to retrieve maximum possible characters. The percentage of occurrence of those characters already calculated earlier using corpus frequency string is added indicating the total percentage retrieval. Eg: retrieved chars: a, r, y and k. If the percentage occurrences of those characters are 8.73, 6.63, 1.24 and 0.58 then the total % retrieval is 8.73 + 6.63 + 1.24 +0.58 = 17.18 as per chart shown in Fig. 1.

English - Probability - Mathcing code points

Sl.N	Plain T	E	T	A	I	N	S	O	R	H	C	D	L	M	U	P	F	G	B	W	Y	V	K	X	J	Q	Z	%
1	1000	E														P										Q		15.70
2	2000	E	T						R	H		D										V	K				Z	37.50
3	4000	E	T	A	I	N			R	H				M	U							V	K				Z	63.97
4	6000	E	T	A	I	N	S	O	R	H				M	U					W		V	K				Z	72.38
5	9000	E			I	N			R	H				M	U					W		V	K				Z	47.13
6	12000	E			I	N			R	H				M	U							V	K			Q	Z	45.99
7	16000	E	T	A	I	N			R	H				M	U							V	K			Q	Z	64.32
8	20000	E	T	A	I	N			R	H					U							V	K			Q	Z	61.20
9	25000	E	T	A	I	N			R	H				M	U							V	K			Q	Z	64.60
10	30000	E	T	A	I	N			R	H				M	U						Y	V	K			Q	Z	66.20
11	40000	E	T	A	I	N			R	H				M	U	P					Y	V	K			Q	Z	68.48
12	50000	E	T	A	I	N			R	H				M	U	P					Y	V	K			Q	Z	63.31
13	70000	E	T	A	I	N			R	H			L	M	U	P	F	G	B	W	Y	V	K			Q	Z	78.27
14	90000	E	T	A	I	N			R	H	C	D	L	M			F	G	B	W	Y	V	K			Q	Z	80.98
15	1E+05	E	T	A	I	N	S	O	R	H	C	D	L	M	U	P	F	G	B	W	Y	V	K	X	J	Q	Z	100.00
		#	#	#	#	13	2	2	14	#	2	3	3	12	#	5	3	3	3	5	6	#	14	1	1	#	14	
		12.32	9.30	8.73	7.86	7.43	6.93	6.92	6.63	4.71	4.27	3.92	3.45	2.58	2.40	2.37	2.12	1.99	1.46	1.29	1.24	0.85	0.58	0.24	0.20	0.14	0.08	

Figure 1 : English probability matching code points

IV. Security Model

a) Normal Method

i. Sample text → encrypted → reverse frequency mapping → decrypted → compared with sample text → calculate % retrieval using corpus frequency string.

b) Proposed Method

ii. Sample text --→ dictionary file + coded file → dictionary file encrypted → reverse frequency mapping → decrypted → compared with sample text → calculate % retrieval using corpus frequency string.

Difference in % retrieval between (a) and (b) is the advantage as per the proposed plan.

Here the text file is converted into a dictionary file which consists of frequency of occurrence of words arranged in descending order and an extended ASCII value is given to each word referred to as the code for that particular word. The extended ASCII values selected are from 33 to 250 i.e. a total of 218 different words are covered. If the number of different words exceed more than 218, then the numbers are repeated appending to its previous values eg 219th word will be 33 33, next will be 33 34 and so on. Based on the coded values of various words in the dictionary, the text file is converted into a coded file which can be decoded only by the dictionary. The dictionary created for different text files is different, hence dynamic in nature. Each text file will have a unique dictionary. The paper [8] speaks about the concept of dictionary and the coded file but left open how the dictionary is to be transmitted in a secure way. It also does not speak of any other language other than English. ie it has not considered language complexity as a factor for security of text data. The coded file created in this paper is different with only the coded values as it serves the purpose. The dictionary is encrypted and transmitted. The coded file is transmitted without encryption. The attacker cannot decode the coded file without getting the information of the dictionary file. The actual percentage of data retrieved from coded file

(which represents the plain text data file) will finally be much less than the percentage retrieved from the dictionary file. The percentage data retrieved from dictionary file in various languages for a fixed corpus sizes have been found out using programs in Python 2.7 and displayed in figures 4 and 5 below.

c) Sample Text

Dance, little baby, dance up high,
Never mind baby, mother is by;
Crow and caper, caper and crow,
There little baby, there you go:
Up to the ceiling, down to the ground,
Backwards and forwards, round and round.
Then dance, little baby, and mother shall sing,
With the merry gay coral, ding, ding, a-ding, ding.

d) Dictionary File

and -> !
baby, -> "
little -> #
the -> $
ding, -> %
to ->&
mother -> '
Never -> (
caper, ->)
dance -> *
is -> +
There -> ,
mind -> -
forwards, -> .
down -> /
high, -> 0
a-ding, -> 1
caper -> 2
Then -> 3
coral, -> 4
there -> 5
Dance, -> 6
ceiling, -> 7
you -> 8
With -> 9
round. -> :
ground, -> ;
Backwards -><
gay -> =
shall ->>
Baby's -> ?
go: -> @
Up -> A
crow, -> B
dance, -> C
merry -> D
The -> E
by; -> F
Crow -> H
ding. -> I
up -> J
round -> K
sing, -> L

e) Coded File

6 # " * J 0(- " ' + F
H !) 2 ! B, # " 5 8 @
A & $ 7 / & $;< ! . K ! :
3 C # " ! ' > L 9 $ D = 4 % % 1 I

% retrieval normal method for sample text is 84.75% and proposed method is 0.0%.

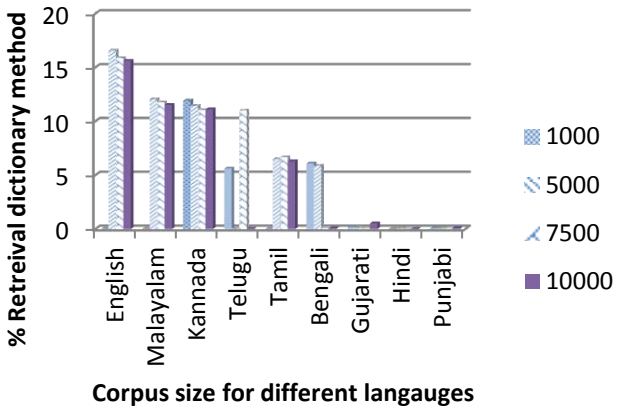


Figure 2 : % Retrieval by dictionary method graph for different languages

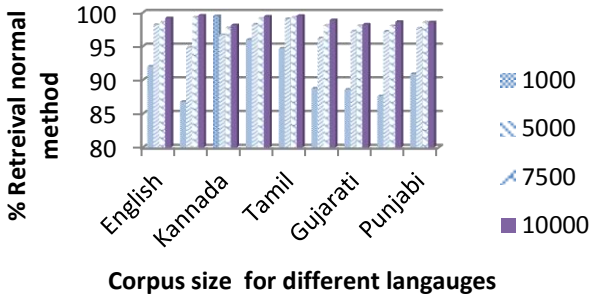


Figure 3 : % Retrieval by normal method graph for different languages

Language	% retrieval dictionary method			
	Corpus size			
	1000	5000	7500	10000
English	0	16.55	15.87	15.6
Malayalam	0	12.03	11.76	11.51
Kannada	11.91	11.43	11.05	11.1
Telugu	5.6	10.78	11.0	11.45
Tamil	0	6.48	6.67	6.27
Bengali	6.07	5.86	0.07	0.05
Gujarati	0.15	0.1	0.05	0.52
Hindi	0	0.02	0.01	0
Punjabi	0.04	0.024	0.045	0.078

Figure 4 : Table displaying the retrieval percentage by dictionary method

Language	% retrieval normal method			
	Corpus size			
	1000	5000	7500	10000
English	92	98.18	98.53	99.18
Malayalam	86.78	94.83	99.31	99.57
Kannada	99.49	96.68	97.74	98.14
Telugu	95.99	98.23	99.11	99.43
Tamil	94.69	99.05	99.27	99.53
Bengali	88.75	96.19	98.09	98.89
Gujarati	88.58	97.25	98.07	98.24
Hindi	87.58	97.19	98.02	98.64
Punjabi	90.89	97.68	98.55	98.57

Figure 5 : Table displaying the retrieval percentage by normal method

V. Findings

a) The percentage data retrieved using conditional probability following a normal method is as explained in IV(a) and after intelligently converting the same sample text into a dictionary form and carrying out the same process as in IV(a) displays a vast difference in the percentage retrieval of data,(Figs 4 & 5) thereby making the proposed system as explained in IV(b) strongly secured.

b) Carrying out the same procedure for similar corpus sizes, the percentage retrieval of data in various Indian languages is far less than English proving that text data transmitted in regional languages is more secured than English language.

c) Amongst the various Indian languages, Gujarati, Hindi and Punjabi display a very low percentage of retrieval of data, making it more secure as far as transmission of data is concerned compared to other languages considered.

d) The three Indian languages Gujarati, Hindi and Punjabi prove to be the most secure amongst the languages considered as case study, are stroke based unlike the languages of the southern part of India which are curvature based.

VI. Conclusions

Security of transmitted data over the internet is most secure when transmitted in any of the Indian languages compared to English language after converting the data into an intermediate form (dictionary and the coded file).

By creating the dictionary, the percentage retrieval compared with plain text file is far less than without creating the dictionary file.

By mapping the retrieved data from dictionary file to coded file the actual data to be retrieved is likely to be far lesser compared to what has been projected in Figs. 4 and 5.

Of the languages considered for case study, Guajarati, Punjabi and Hindi provide better security and they happen to be stroke based than curvature based (south Indian languages).

References Références Referencias

1. Jakobsen, T: A fast Method for Cryptanalysis of Subsittution Ciphers. J. Cryptologia, Volume 19, Issue 3 1995, pp. 265-274.
2. Adam Stone: Internationalizing the Internet. J. Internet Computing. 3, 2003, pp. 11-12.
3. Bauer F L: Decrypted secrets-Methods and Maxims of Cryptology, Springer, 2007.
4. Menezes A. J. P: Handbook of Applied Cryptography. CRC Press, 2001.
5. Lee K.W., C.E. Teh, Y.L. Ta: Decrypting English Text Using Enhanced Frequency Analysis: National Seminar on Science, Technology and Social Sciences 2006 (Ui TM-STSS 2006). pp. 1-7.
6. Bhadri Raju MSVS, Vishnu Vardhan B, Naidu G A, Pratap Reddy L, Vinaya Babu A : Effect of Language Complexity on Deciphering Substitution Ciphers - A Case Study on Telugu.
7. Devasish Pal, Raju Ejjagiri, Dr. A Vinaya Babu: Complexity of Bengali Language and its relation to data security volume1 Issue 4 - 2012 (IJACIT) ISSN 2277-9140.
8. Dr. V.K. Govindan, B.S. Shajee mohan:An Intelligent text data encryption and compression for high speed and secure data transmission over internet.
9. Bao-Chyuan Guan, Ray-I Chang, Yung Chung Wei, ChiaLing Hu, Yu-Lin Chiu: An encryption scheme for largeChinese texts: IEEE 37th Annual.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Survey on Efficient Audit Service to Ensure Data Integrity in Cloud Environment

By Jaspreet Kaur & Jasmeet Singh
Punjab Techincal University

Abstract - Cloud computing is an internet based computing which provides different users an opportunity to store their data in the cloud. While data outsourcing relieves the owner of the burden of the local data storage and maintenance but as they have no longer physical possession of outsourced data makes data integrity protection a very challenging task. This paper explores the secure cryptographic hash function along with some other techniques that can be used by TPA to ensure the integrity of data stored in the cloud at regular intervals or on user request.

Keywords : *cloud, data integrity, secure hash algorithm, station-to-station protocol, third party auditor.*

GJCST-C Classification: *H.2.7*

Strictly as per the compliance and regulations of:

Survey on Efficient Audit Service to Ensure Data Integrity in Cloud Environment

Jaspreet Kaur [α] & Jasmeet Singh [σ]

Abstract - **Cloud computing is an internet based computing which provides different users an opportunity to store their data in the cloud. While data outsourcing relieves the owner of the burden of the local data storage and maintenance but as they have no longer physical possession of outsourced data makes data integrity protection a very challenging task. This paper explores the secure cryptographic hash function along with some other techniques that can be used by TPA to ensure the integrity of data stored in the cloud at regular intervals or on user request.**

Keywords : cloud, data integrity, secure hash algorithm, station-to-station protocol, third party auditor.

I. Introduction

Cloud computing is a general term for anything that involves delivering hosted services over the internet. These services are broadly divided into three categories:

a) SaaS (Software as a Service)

SaaS is a model of software deployment where consumer use the provider's application running on a cloud infrastructure through a thin client interface.

b) PaaS (Platform as a Service)

The platforms used to develop, build and test applications are provided to the consumer by the cloud.

c) IaaS (Infrastructure as a Service)

This is a model in which service provider owns the equipments used to support operations, including storage, hardware, servers and networking components. The client typically pays on a per-use basis. [3]

Cloud computing is becoming more and more popular now a days, where data is outsourced into the cloud. Its pros include relief of the burden of the storage management, universal data access with independent geographical locations, and avoidance of capital expenditure on hardware, software and personnel maintenance. However, outsourcing data introduces new security issues. [4]

The first issue is data integrity. In computer security, data integrity can be defined as "the state that exists when computerized data is the same as that in the source document and has not been exposed to accidental or malicious alterations or destruction". Integrity of data stored at the entrusted cloud server is not guaranteed. [1, 4]

The second issue is unfaithful cloud server providers (CSP). There are many reasons why CSPs are not always trustworthy like, for saving money and storage space, CSPs may discard the data that has not been accessed for long time (which belongs to ordinary client) or sometimes even hide data losses or corruptions to maintain a reputation.

As data owners outsourced their data to the cloud and do not maintain the local copy, so simple cryptographic measures cannot be used directly to monitor the integrity of data. Also simply downloading the data for monitoring integrity is not a viable solution as it incurs high cost of input/output and transmission across the network.

To check the integrity of data only while accessing is not sufficient as the un-accessed data left unchecked from the verification process and it might get too late to recover any loss or damage to the unchecked data.

In addition, from the system usability point of view, data owners should be able to just use cloud storage as if it is local, without worrying about the need to verify the correctness of data. Therefore, an external third party auditor (TPA) is required. [4]

The TPA is an independent authority that has expertise and capabilities to monitor the integrity of cloud data outsourced by the client and informs him about data corruption or loss, if any.[6] To securely introduce an effective third party auditor (TPA), the following two fundamental requirements have to be met:

1. TPA should be able to audit the cloud data storage efficiently without asking for the local copy of data thereby reducing the on-line burden of cloud users.
2. The third party auditing process should not affect user data privacy.[7]

Figure 1 represents the cloud storage architecture which consists of three different entities: Users, Cloud Storage Server and Third Party Auditor (TPA).

Author α : M.TECH (Computer Science and Engineering) RIMT-IET, Mandi Gobindgarh, Punjab (India). E-mail : kkhushi12@yahoo.co.in
Author σ : Asst. Professor (CSE Department) RIMT-IET, Mandi Gobindgarh, Punjab (India). E-mail : jasmeetgurm@gmail.com

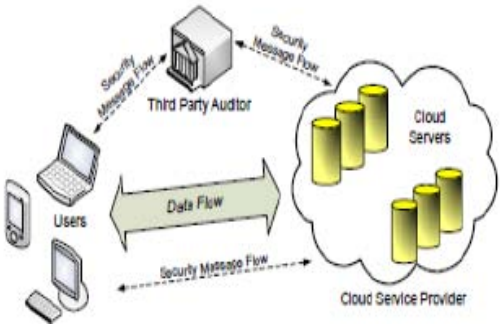


Figure 1 : Architecture of cloud data storage service

1. *User*

Users are the data owners who have the large amount of data to be stored in the cloud and access them when needed. Users depend on cloud storage server for data computation.

2. *Cloud Storage Server*

A cloud storage server (CSS) is an entity that is managed by cloud service provider (CSP). It provides space to the user for data storage and computation.

3. *Third Party Auditor (TPA)*

An TPA is an entity who has capabilities to verify the integrity of cloud data on behalf of the user's request. [1]

II. Related Work

In [9] author implemented mechanism in which integrity is checked at 2 sides-by cloud server (for inside attack) and by TPA (for outside attack) using digital signature with MD5.But we analyze that as cloud server is not trustworthy so data should not be exposed to the cloud. Instead user can rely on TPA who has expertise in verification process and moreover user can authenticate TPA to check for its credibility.

In [5] author proposed auditing scheme in which TPA checks the integrity of data outsourced by user in the cloud. For monitoring integrity auditor takes cipher text from the cloud and generates original data from it using XOR. TPA calculates the hash value using SHA-1 algorithm and compares it with the hash value taken from user. If the value matches it is ensured that data is safe, otherwise tampered.

This approach relieves the user from worry of downloading data and verifying its correctness. This also preserves user resources that could be consumed otherwise. But there is lack of entity authentication between user and TPA which is necessary to trust third party. So we analyze that it could be better to use modified version of Diffie-Hellman i.e. STS (Station-To-Station) protocol in place of Diffie-Hellman which provides entity authentication along with key generation. Because of the security flaws (Collision Attacks) found in SHA-1 it is preferable to use SHA-2 which is more secure and strong.

III. Techniques

a) Station-to-Station protocol(STS)

In public-key cryptography, the Station-to-Station (STS) protocol is a cryptographic key agreement scheme consists of Diffie-Hellman key establishment followed by an exchange of authentication signatures. In this protocol, we assume that the parameters used for the key establishment are fixed and known to all users.

i. *STS Setup*

The following data must be generated before initiating the protocol:-

- An asymmetric signature key pair for each party-Required for authentication. The public portion of this key pair may be shared prior to session establishment.
- Key establishment parameter-The specification of a particular cyclic group and the corresponding primitive element α. These parameters may be public.

Sharing this data prior to the beginning of the session lessens the complexity of the protocol.

ii. *Basic STS*

Supposing all setup data has been shared, the STS protocol proceeds as follows: (All exponentials are in the group specified by p)

- Alice generates a random number x and computes the exponential α^x and send it to Bob.
- Bob generates a random number y and computes the exponential α^y.
- Bob computes the shared secret key $K = (\alpha^x)^y$.
- Bob concatenates the exponentials (α^y, α^x) (order is important), signs them using his asymmetric key B, and then encrypts them with K. He sends the cipher text along with his own exponential α^y to Alice.
- Alice computes the shared secret key $K = (\alpha^y)^x$.
- Alice decrypts and verifies Bob's signature.
- Alice concatenates the exponentials (α^x, α^y) (order is important), signs them using her asymmetric key A, and then encrypts them with K. She sends the cipher text to Bob.
- Bob decrypts and verifies Alice's signature.

Alice and Bob are now mutually authenticated and have a shared secret. This secret, K, can then be used to encrypt further communication. The basic form of the protocol is formalized in the following three steps as shown in figure 2. [11]

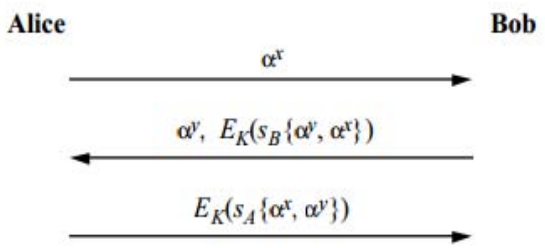


Figure 2 : Station to Station Protocol Steps

b) Exclusive or Operation

XOR is an operation in which same bits produce a resultant bit as 0 whereas different bits produce a resultant 1. In this, a XOR operation is done at original text along with secret value which gives cipher text. The original text can be obtained by performing an XOR operation between the secret key and resultant cipher text.

c) SHA-2

SHA stands for Secure Hash Algorithm. SHA-2 is the collective name of one-way hash functions developed by the NIST. SHA-256, SHA-384, and SHA-512 pertain to hashes whose outputs are 256 bits, 384 bits and 512 bits, respectively.

A hash function is an algorithm that transforms (hashes) an arbitrary set of data elements into a single fixed length value (the hash). The computed hash value may then be used to check the integrity of copies of the original data without providing any means to derive the source (irreversibly). A hash value therefore may be freely distributed or stored as it is only used for comparative purposes. SHA-2 features a higher level of security than its predecessor, SHA-1.

The comparisons between the SHA parameters are shown in table 1.

	sha-1	sha-256	sha-384	sha-512
message digest size	160	256	384	512
message size	$<2^{64}$	$<2^{64}$	$<2^{128}$	$<2^{128}$
block size	512	512	1024	1024
word size	32	32	64	64
number of steps	80	64	80	80
security	80	128	192	256
notes : 1.all sizes are measured in bits 2. security refers to the fact that a birthday attack on a message digest of size n produces a collision with a work factor of approximately $2^{n/2}$				

Table 1 : Comparison of SHA Parameters

The security provided by a hashing algorithm is entirely dependent upon its ability to produce a unique value for any specific set of data. When a hash function produces the same hash value for two different sets of data then a collision is said to occur. Collision raises the possibility that an attacker may be able to computationally craft sets of data which provide access to information secured by the hashed values of pass codes or to alter computer data files in a fashion that would not change the resulting hash value and would thereby escape detection. A strong hash function (e.g. SHA-2) is one that is resistant to such computational attacks. [10]

An overview of SHA-256 is given here, and then the differences between SHA-256 and the other members of the SHA-2 family are outlined. The SHA-256 algorithm essentially consists of 3 stages: (1) message padding and parsing; (2) expansion; and (3) compression.

i. *Message Padding and Parsing*

The binary message to be processed is appended with a '1' and padded with zeros until its length 448 mod 512. The original message length is then appended as a 64-bit binary number. The resultant padded message is parsed into N 512-bit blocks, denoted $M^{(1)}$, $M^{(2)}$, ……. $M^{(N)}$. These $M^{(i)}$ message blocks are passed individually to the message expander.

ii. *Message Expansion*

The functions in the SHA-256 algorithm operate on 32-bit words, so each 512-bit $M^{(i)}$ block from the padding stage is viewed as 16 32-bit blocks denoted $M_t^{(i)}$, $0 \leq t \leq 15$.The message expander (also called the message scheduler) takes each $M^{(i)}$ and expands it into 64 32-bit W_t blocks.

iii. *Message Compression*

The W_t words from the message expansion stage are then passed to the SHA compression function, or the 'SHA core'. The core utilizes 8 32-bit working variables labeled A, B, . . . , H, which are initialized to predefined values $H_0^{(0)}$–$H_7^{(0)}$ at the start of each call to the hash function. Sixty-four iterations of the compression function are then performed and intermediate hash value $H^{(i)}$ is calculated:

$$H_0^{(i)} = A + H_0^{(i-1)},\ H_1^{(i)} = B + H_1^{(i-1)},\ \ldots,\ H_7^{(i)}$$

The SHA-256 compression algorithm then repeats and begins processing another 512-bit block from the message padder. After all N data blocks have been processed, the final 256-bit output, $H^{(N)}$, is formed by concatenating the final hash values:

$$H^{(N)} = H_0^{(N)} \,\&\, H_1^{(N)} \,\&\, H_2^{(N)} \,\&\, \ldots \,\&\, H_7^{(N)}$$

iv. *SHA-2 Algorithm Differences*

The SHA-512 algorithm has a similar structure to the SHA-256 algorithm, where: (i) it processes messages in blocks of 1024 bits rather than 512 bits; (ii) it uses 64-bit operations instead of 32-bit operations; (iii) it iterates its compression function 80 times rather than 64 times, but are otherwise similar in structure.[8]

IV. Conclusions

Cloud Computing today is the beginning of network based computing over internet in force. So monitoring integrity of cloud data storage is of critical importance. For this, third party auditor who has expertise in verification process can monitor integrity on behalf of user. The techniques mentioned above can be used to achieve the auditing task efficiently. We believe that monitoring integrity of cloud storage data is very much needed as data in cloud is not secure.

References Références Referencias

1. Reenu Sara Georeg, Sabitha S," Survey on Data Integrity in Cloud Computing", International Journal

of Advanced Research in Computer Engineering & Technology (IJARCET) Vol. 2, Issue 1, January 2013.

2. K. Govinda, V. Gurunathprasad, H. Sathishkumar, "Third Party Auditing for Secure Data Storage in Cloud through Digital Signature using RSA", International Journal of Advanced Scientific and Technical Research", Vol. 4, Issue 2, August 2012.
3. Akkala Saibabu, T. Satyanarayana Murthy, "Security Provision in Publicly Auditable Secure Cloud Data Storage Services using SHA-1 Algorithm", International Journal of Computer Science and Information Technologies (IJCSIT) Vol. 3(3), 2012.
4. Changsheng Wan, Juan Zhang, Zhongyuan Qin, "A XOR based Public Auditing Scheme for Proof-of-Storage".
5. K. Govinda, E. Sathiyamoorthy, "Data Auditing in Cloud Environment using Message Authentication Code", International Conference on Emerging Trends on Advanced Engineering Research (ICETT), 2012.
6. Muralikrishnan Ramane, Bharath Elangovan, "A MetaData Verification Scheme for Data Auditing in Cloud Environment", International Journal on Cloud Computing: Services and Architecture (IJCCSA), Vol.2, No.4, August 2012.
7. Lingaraj Dhabale, Priti Pavale, "Providing Secured Data Storage by Privacy and Third Party Auditing in Cloud", International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.
8. Robert P. McEvoy, Francis M. Crowe, Colin C. Murphy and William P. Marnane, "Optimisation of the SHA-2 Family of Hash Functions on FPGAs".
9. Dalia Attas, Omar Batrafi," Efficient integrity checking technique for securing client data in cloud computing", International Journal of Electrical & Computer Sciences (IJECS-IJENS), Vol: 11, No: 05.
10. http://en.wikipedia.org/wiki/SHA-2
11. http://en.wikipedia.org/wiki/Station-to-Station_protocol

Global Journals Inc. (US) Guidelines Handbook 2013

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

- 'FARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'FARSC" can be added to name in the following manner. eg. **Dr. John E. Hall, Ph.D., FARSC or William Walldroff Ph. D., M.S., FARSC**
- Being FARSC is a respectful honor. It authenticates your research activities. After becoming FARSC, you can use 'FARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 60% Discount will be provided to FARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- FARSC will be given a renowned, secure, free professional email address with 100 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- FARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 15% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- Eg. If we had taken 420 USD from author, we can send 63 USD to your account.
- FARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- After you are FARSC. You can send us scanned copy of all of your documents. We will verify, grade and certify them within a month. It will be based on your academic records, quality of research papers published by you, and 50 more criteria. This is beneficial for your job interviews as recruiting organization need not just rely on you for authenticity and your unknown qualities, you would have authentic ranks of all of your documents. Our scale is unique worldwide.
- FARSC member can proceed to get benefits of free research podcasting in Global Research Radio with their research documents, slides and online movies.
- After your publication anywhere in the world, you can upload you research paper with your recorded voice or you can use our professional RJs to record your paper their voice. We can also stream your conference videos and display your slides online.
- FARSC will be eligible for free application of Standardization of their Researches by Open Scientific Standards. Standardization is next step and level after publishing in a journal. A team of research and professional will work with you to take your research to its next level, which is worldwide open standardization.

- FARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), FARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 80% of its earning by Global Journals Inc. (US) will be transferred to FARSC member's bank account after certain threshold balance. There is no time limit for collection. FARSC member can decide its price and we can help in decision.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

- 'MARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'MARSC" can be added to name in the following manner. eg. Dr. John E. Hall, Ph.D., MARSC or William Walldroff Ph. D., M.S., MARSC
- Being MARSC is a respectful honor. It authenticates your research activities. After becoming MARSC, you can use 'MARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 40% Discount will be provided to MARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- MARSC will be given a renowned, secure, free professional email address with 30 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- MARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 10% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- MARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- MARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), MARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 40% of its earning by Global Journals Inc. (US) will be transferred to MARSC member's bank account after certain threshold balance. There is no time limit for collection. MARSC member can decide its price and we can help in decision.

AUXILIARY MEMBERSHIPS

ANNUAL MEMBER

- Annual Member will be authorized to receive e-Journal GJCST for one year (subscription for one year).
- The member will be allotted free 1 GB Web-space along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted free 500 MB email space.

PAPER PUBLICATION

- The members can publish paper once. The paper will be sent to two-peer reviewer. The paper will be published after the acceptance of peer reviewers and Editorial Board.

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.Online Submission: There are three ways to submit your paper:

(A) (I) 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 Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

INDEX

P

R

S

T

U

V

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

ISSN 9754350