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

Global Journal

of Computer Science and Technology: C

Software & Data Engineering

Quality of Service Centric Improved Approaches to Handle

Highlights

Assessing Composition Impact

Data Mining in Biodata Analysis

Discovering Thoughts, Inventing Future

VOLUME 14 ISSUE 9 VERSION 1.0

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

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

Volume 14 Issue 9 (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 Headquarters
301st Edgewater Place Suite, 100 Edgewater Dr.-Pl,
Wakefield MASSACHUSETTS, Pin: 01880,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

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

Packaging & Continental Dispatching

Global Journals
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin:452009, India

Find a correspondence nodal officer near you

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

eContacts

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

Pricing (Including by Air Parcel Charges):

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

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

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 Issue

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 9 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Digital Data Theft Detection using Watermarking

By B. Sai Sindhush, R. V Keshava Rao & Dr R. Bulli Babu

KL University, India

Abstract- Large amount of data is embedded in media and spread in the internet. This data can be replaced easily with the help of some software. Digital watermarking is a very useful technology in today's world, to prevent illegal copying of data. Digital watermarking can be applied to all forms of multimedia.

Keywords: copyright protection, digital watermarking, steganography, information hiding, robustness.

GJCST-C Classification : H.2.7

Strictly as per the compliance and regulations of:

Digital Data Theft Detection using Watermarking

B. SaiSindhush [α], R.V KeshavaRao [σ] & Dr R. BulliBabu [ρ]

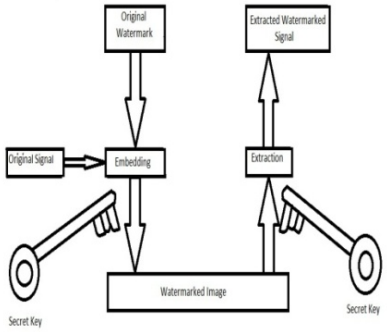
Abstract- **Large amount of data is embedded in media and spread in the internet. This data can be replaced easily with the help of some software. Digital watermarking is a very useful technology in today's world, to prevent illegal copying of data. Digital watermarking can be applied to all forms of multimedia.**

Keywords: copyright protection, digital watermarking, steganography, information hiding, robustness.

I. Introduction

In computer science information hiding or hiding data in a message is the important principle of steganography.Information hiding is mainly divided into three categories Cryptography, Steganography, and Watermark. Cryptography is the process of converting comprehensible data into unintelligible data that can't be able understand by unauthorized people. The authorized user with the key can decrypt the ciphertext. As many modification were made in the field of multimedia and communications, now it became easy for the unauthorized users to decrypt a ciphertext into comprehensible data. Hence more complicated methods were developed to provide higher security than cryptography. These techniques are known as Steganography and Watermarking. Steganography is the time taking process. It hides data over a cover object in such a way that the sence of data is not detected by the hacker. Watermarking is related to the steganography. There is one main point in watermarking is that the invisible data is related to the cover object. Watermarking is mainly used for copyright protection, user authentication and security. Digital watermarking is the process of embedding a digital signal (audio, video or image) or hide a small digital data in comprehensible data which can not be easily removed is called digital watermarking. Digital watermarking is also called data hiding.

Watermarking block diagram



watermarking system is divided into three types embedding[1], attack and detection. In embedding technique an algorithm accept user and data as input to be embedded and implement the watermark signal. Then watermark signal is send to another host. If this person makes any changes to the watermark signal is called an attacking. There are various types of attack is possible on the watermarked signal. Detection is an algorithm which takes attacked data as input and Extract the watermark data form the attacked data

II. Types of Digital Watermaking

There are two types of digital watermarking, they are

a. Visible watermark
b. Invisible watermark

a. *Visible watermark*- Visible watermark contains visible data or a band logo, used for the owners identification.In visible watermarking, the watermark signal is visible in the picture, video or text.

Example- Logo of the channels such as Animal planet, SONY..etc is on the right top corner of the television, it is visible

Simple watermarked image

Author α σ: B Tech, Department of Electronics and Computer Engineering KL University,India. e-mail: keshav372@gmail.com
Author ρ: Associate Professor, Department of Electronics and Computer Engineering KL University,India.

b. *Invisible Watermark*- In invisible watermark the watermark data is not visible to user. The watermark is encoded in such a way that the watermark data is not visible to the unauthorized users (Attacker)[2]. Inivisiblewayermarking is also used for the purposeof image identification and provide security to the image from being used by unauthorized users. Invisible watermarking also contains of encode and decode process.

Watermark insertion is represented as: O" = EU (A,W)

Where O is the original image, W is the watermark information being embedded, U is the user's insertion key, and Erepresents the watermark insertion function.

Invisible Watermarking[1] (Least significant bit watermarking)- Least significant bit watermarking is themost secured technique of watermarking. It also be applied to both visible and invisible watermarking. Spatial domain technique changes the pixels of one or two subset of the image.

Let us see the one example on image watermarking process

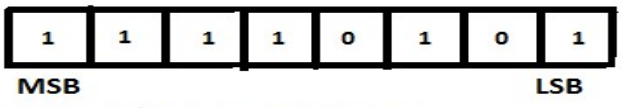
Steps-

1. For the image testing standard images A and B will be selected. The base image or original image is A for which watermaking is added .the watermarking image is B that will be added to the original image A.
2. The least significant bits(LSB) of the original image A will replaced with the most significant bits(MSB) of the water marking image B[2]
3. The resulting image that comes after the combination of both A and B images is Final image C will be watermarked image.

Hence C contains an image A which LSB bits are replaced with the MSB of the image B. The original image and watermarking image is taken in binary code form-

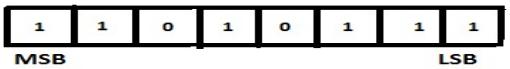
WatermarkImage=11110101

Base Image = 11010111

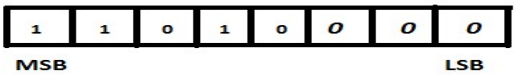


It is 8 bit image. In this case consider bits= 3
Therefore whole frame is moved (8-3= 5) by 5 placed to the right, thereby passing the MSB to the LSB.

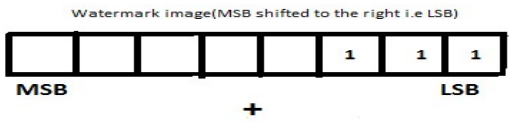
Base image=

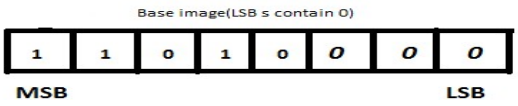


From the base imge, the LSB s(last three bits of base image) are set to O

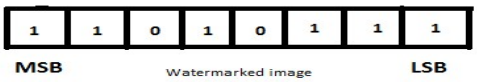


Here the LSB of the original image is replaced with zeros and the MSB bits of the watermark image will be shifted to LSB bits





Final image=



The Final water marked image contains 5 MSB bits of orginal image and 3LSB bits of the water mark image

Base Image

Watermark Image

watermarked Image

III. Requirements of Digital Watermarking

The requirements of digital watermarking are

A. Transparency- making the watermark image clear and transparent without effecting the quality of original image .

B. Robustness- This is one of the requirements of the watermarking .it means the watermark which is designed must be resistible to all kinds of attacks by the unauthorized users and hackers.

C. Capacity- it describes the amount of data that can be embedded into multimedia formats such as image, audio, video or text for retrieving the prefect data of watermark during extraction.

IV. Conclusion

In this paper we describes about different types of watermarking and its techniques. There are two types of digital watermarking techniques they are visible and invisible watermarking techniques. It provides authentication for owners. hence by using this watermaking techniques the data can be protected and stored from the unauthorized users.

V. Acknowledgement

We would like to give thanks toDr.R.Bullibabu for his guidance and help us to complete this paper.

References Références Referencias

1. I.J . Cox et al , "Digital Watermarking and Steganography" (Second edition), Morgan Kaufmann, 2008.
2. W. Bender D. Gruhl N. Moromoto and A. LU Techniques for data hiding. IBM Systems Journals, 35(3 -.

Global Journal of Computer Science and Technology (C) Volume XIV Issue IX Version I Year 2014

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 9 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Data Mining in Biodata Analysis

By D. Aruna Kumari D. Poojitha Bhavana & V. Venkata Sai Aditya

KLEF University,India

Abstract- For finding interesting patterns in large databases has lot of development in recent years.. Data mining is used in many fields like medicine, securing the data etc. Whereas bio data means the data regarding the biology, medical science, DNA technology and Bioinformatics in-depth analysis. Bio Informatics is the science which can perform managing, finding data, integrating, interrupting information from biological data, genomic, and metadata. Even additional knowledge and complexness can lead to the integration among genes. This paper is all about joining these two fields, the data regarding biology us ing data mining and gives the details of future developments in biodata analysis.

GJCST-C Classification : *H.2.8*

Strictly as per the compliance and regulations of:

Data Mining in Biodata Analysis

D. Aruna Kumari [α] D. Poojitha Bhavana [σ] & V. Venkata Sai Aditya [ρ]

Abstract- For finding interesting patterns in large databases has lot of development in recent years.. Data mining is used in many fields like medicine, securing the data etc. Whereas bio data means the data regarding the biology, medical science, DNA technology and Bioinformatics in-depth analysis. Bio Informatics is the science which can perform managing, finding data, integrating, interrupting information from biological data, genomic, and metadata. Even additional knowledge and complexness can lead to the integration among genes. This paper is all about joining these two fields, the data regarding biology us ing data mining and gives the details of future developments in biodata analysis.

I. Introduction

There are distinct changes in medical research and biodata analysis and there is a lot of growth in medical data collected in medical studies and cancer therapy studies by inventing sequencing patterns, protein-protein interactions gene functions. In biotechnology and bio-data analysis there is a fast growth which has led to the rapid growth in new fields like biodata analysis.

At the same time, according to the recent progress there is a lot of development in the methods of mining interesting patterns and information in large databases, starting from efficient classification methods to clustering, frequent, , serial and structured pattern analysis methods, outlier analysis and visualization.

This paper is about how to combine these two fields i.e. data mining and biodata analysis. We need to analyze in which way data mining is helpful in biodata analysis and overview few research problems that may analyze further developments.

a) Themes of Biodata Analysis

i. Data Cleaning, Data Pre-Processing and Data Integration

By applying several techniques a variety of bio medical sciences are in use with different geographical dimensions. These are based on data values in bio medical information, genome or proteome databases.Data should be gathered, characterized and clean to extract and analyse information from medicine database and heterogeneous database. The steps for this processing are time taking factors. They need multiple scans for enormous databases to ensure the standards, as a result of he terogeneous and distributed nature of data there are many challenges in the analysis of medical data. Data cleaning, Data pre-processing and data integration helps in the integration of biomedical data and in the formation of data warehouses for biomedical analysis.

ii. Exploring of Existing Data Mining Tools for Bio Data Analysis

Due to a lot of development in data mining, there are several data mining, machine learning, and applied mathematical analysis systems and tools offered for general data analysis. This analysis is often utilized in biodata analysis and exploration. Data mining analysis is used for biodata analysis including SAS enterprise miner, IBM Intelligent Miner, Microsoft SQLServer 2000, SGI MineSet, and InxightVizServer.Biospecific data analysis systems like GeneSpring, Spot Fire, and VectorNTI can be used in biodata analysis. There are different types of software tools that are developed for resolving the basic bio medical issues. These tools are developing fastly as well..ForBiodata analysis researches should be well trained regarding the usage of tools.

There is much scope for researchers for data mining methods in biodata analysis. Some topics in this view are as follows:

b) Similarity in Search and Comparison in Biodata

An essential problem in Biodata analysis is searching similarly and comparing the bio-sequential structures supporting their essential options and functions. For example, the sequences of the genes which are unhealthy and healthy will be compared to notice to note the distinction between the two varieties of genes. This can be achieved by taking the two categories of genes, then finding the king of factor whether the gene is unhealthy or the healthy one, then comparing the more oftenly occurring patterns of every class. Generally the genetic factor of the disease can be indicated in a way that the diseased sample patterns occur more ofenlyoccuring than the healthy sample patterns. The sequences occurring more frequently in healthy samples indicates the mechanism that protects the body from the diseases. Same type of research can be done on microarray data and protein data to spot the differences in the patterns. Moreover, as the biodata sometimes contains non-perfect matches, it is necessary to develop sequential pattern mining algorithms with in the noisy environment.

Author α σ ρ: Department of Electronics and Computer engineering KLEF University Vaddeswaram, Tadepalli(Mandal), Guntur(dist), 522502 A.P, INDIA. e-mails: Aruna_D@kluniversity.in, poojitha.daita@gmail.com, saiaditya11994@gmail.com

c) Association Analysis

There are a lot of studies which has concentrated on the comparing one gene with the other gene. But most of the diseases are no occurred just only by one gene ,it may occur by the combination of two or more genes.Association and correlation analysis strategies can be used to determine the types of genes that may cooccur in final samples. Discovery of groups of such genes or proteins can be done by such analysis. Study of interactions and relationships among the groups and protiens can be done with the help of the analysis done by the association analysis.It is important to develop the serial or structural pattern mining algorithms in the mining environment because the biodata information usually consists of noise or non-perfect matches.

d) Cluster Analysis

The process of grouping a group of objects into clusters in which there are similarities in the objects in the same cluster is high and in the objects of different clusters is low is called as clustering.. Clustering is not only in pattern recognition, marketing, social and scientific studies but also in Biodataanalysis.Either Euclidean distances or density are used to determine the algorithms of cluster analysis. The features of biodata analysis are high dimension space, and it is troublesome to review the differentials with scaling and shifting factors in multi-dimensional space and discover the frequently occuring patterns.

e) Path Analysis

Complex network among the genes is formed by the biological process. These networks are build ,modeled and visualized using path analysis. The information about biochemical reactions is stored in the database by using the pathway tools. A single genes may not be the reason for causing the disesase,it may be a group of genes responsible for causing a disease process.At the same time there are different stages for different diseases which may become active in any stage of the disease process. The stages of the disease development process will be having a sequence of genetic activities. When this sequence is recognized it will be easy to find the type of the disease for which the future researches can also be developed. By this we can give a better treatment to the diseased people.

f) Data Visualization and Visual Data Mining

For aiding the data comprehension the capabilities of human visual systems is used with the help of computer generated representations.. AVS, SGI Explorer, Khoros, MatLab, Visage, SPSS are the general visualization software products. There are many factors for visual data mining and data visualization in the biomedical domain. The first is its huge size.It creates complexities and diversity in biomedical databases. Second, the data producing biotechnologies have been processing rapidly. The demand for biomedical services has been rapidly increasing. Serial patterns of genes are represented by using Graphs, trees, cubes, and chains by different visualization tools.

g) Privacy Preserving Mining of Bio-Medical Data

Privacy preserving is the most important factor that any field should have .In biomedical data analysis data regarding genes, proteins, research details should be maintained carefully. For this purpose privacy preserving technique is used. Authorities of hospitals and research institutes will not be able to give the information regarding their hospital details, patient details, their research details etc. Everything should be maintained secretly. Moreover giving such details to other is a crime. So all the details should be secretly maintained. For this purpose privacy preserving should be done with the help of datamining methods which preserves the biomedical data.

II. Conclusion

The research frontiers which are data mining and bioinformatics are fast expanding. The research issues in bioinformatics should be examined and the new data mining methods are developed for biodataanalysis which are effective and scalable. There are many methods in data mining which can be used in any field .In biodata analysis data can be preserved, compared, similarities can also be checked using data mining.

References Références Referencias

1. R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD'00,* pp. 439–450, Dallas, TX, May 2000.
2. A. Baxevanis and B. F. F. Ouellette. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (2nd ed.).* John Wiley & Sons, 2001.
3. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, New York, 2001.
4. H. Wang, J. Yang, W. Wang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *SIGMOD'02,* pp. 418–427, Madison, WI, June 2002.
5. J. Yang, P. S. Yu, W. Wang, and J. Han. Mining long sequential patterns in a noisy environment. In *SIGMOD'02,* pp. 406–417, Madison, WI, June 2002.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 9 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Improved Approaches to Handle Bigdata through Hadoop

By K. Sandeep, K. Kondaiah, A. ineetha & Ch. Monica

KLEF University, India

Abstract- Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.Today's world produces a large amount of data from various sources, records and from different fields termed as "BIG DATA". Such huge data is to be analyzed, and filtered using various techniques and algorithms to extract the interested and useful data to gain knowledge. In the new era with the boom of both structured and unstructured types of data, in the field of genomics, meteorology, biology, environmental research and many others, it has become difficult to process, manage and analyze patterns using traditional databases and architectures. It requires new technologies and skills to analyze the flow of material and draw conclusions. So, a proper architecture should be understood to gain knowledge about the Big Data. The analysis of Big Data involves multiple distinct phases such as collection, extraction, cleaning, analysis and retrieval.

GJCST-C Classification : *H.2.8, H.2.6*

Strictly as per the compliance and regulations of:

Improved Approaches to Handle Bigdata through Hadoop

K. Sandeep [α], K. Kondaiah [σ], A. vineetha [ρ] & Ch. Monica[ω]

Abstract- Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.Today's world produces a large amount of data from various sources, records and from different fields termed as "BIG DATA". Such huge data is to be analyzed, and filtered using various techniques and algorithms to extract the interested and useful data to gain knowledge. In the new era with the boom of both structured and unstructured types of data, in the field of genomics, meteorology, biology, environmental research and many others, it has become difficult to process, manage and analyze patterns using traditional databases and architectures. It requires new technologies and skills to analyze the flow of material and draw conclusions. So, a proper architecture should be understood to gain knowledge about the Big Data. The analysis of Big Data involves multiple distinct phases such as collection, extraction, cleaning, analysis and retrieval.This paper presents detailed analysis of Hadoop and MapReduce programming Model and also the challenges that Apache Hadoop, the popular data storage and analysis platform used by major number of large companies is facing and future scope of implementation of Hadoop and various other new improvements to the challenges.

I. Introduction

Apache Hadoop, the popular data storage and analysis platform, has generated a great deal of interest recently. Large and successful companies are using it to do powerful analyses of the data they collect. Hadoop offers two important services: It can store any kind of data from any source, inexpensively and at very large scale, and it can do very sophisticated analysis of that data easily and quickly.

Unlike older database and data warehousing systems, Hadoop is different and those differences can be confusing to users. What data belongs in a Hadoop cluster? What kind of questions can the system answer? Understanding how to take advantage of Hadoop requires a deeper knowledge of how others have applied it to real-world problems that they face.

This paper presents detailed analysis of Hadoop and MapReduce programming Model and also the challenges that hadoop is facing and future scope of implementation of hadoop and various other new algorithms.

Author α σ ρ ω: Department of Electronics and Computer engineering KLEF University. Vaddeswaram, Tadepalli (Mandal), Guntur(dist), 522502 A.P, INDIA. e-mails: Sandeep.k@gmail.com, Kondaya.Kuppala@gmail.com, mailvivacious@gmail.com, monica cherukuri09@gmail.com

II. What is Hadoop?

Hadoop is data storage and processing system. It is scalable, fault-tolerant and distributed. Hadoop was originally developed by the world's largest internet companies to capture and analyze the data that they generate. Unlike older platforms, Hadoop is able to store any kind of data in its native format and to perform a wide variety of analyses and transformations on that data. Hadoop stores terabytes, and even petabytes, of data inexpensively. It is robust and reliable and handles hardware and system failures automatically, without losing data or interrupting data analyses. Hadoop runs on clusters of commodity servers. Each of those servers has local CPU and storage. Each can store a few terabytes of data on its local disk.

Hadoop supports applications under a free license. Three critical components of Hadoop system are:

1. Hadoop Common : Common Utilities Package
2. HFDS: Hadoop Distributed File System with high throughput access to application data.
3. MapReduce: A software framework for distributed processing of large data sets on computer clusters.

*The Hadoop Distributed File System, or HDFS:*HDFS is the storage system for a Hadoop cluster. When data arrives at the cluster, the HDFS software breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.

A distributed data processing framework called Map Reduce: Because Hadoopstores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer. MapReduce isthe plumbing that distributes the work and collects the results.

Hadoop is high-performance distributed data storage and processing system. Its two major subsystems are HDFS, for storage, and MapReduce, for

parallel data processing.Hadoop automatically detects and recovers from hardware and software failures. HDFS and MapReduce will help in performing this.

Hadoop stores any type of data, structured or complex, from any number of sources, in its natural format. No conversion or translation is required on ingest. Data from many sources can be combined and processed in very powerful ways, so that Hadoop can do deeper analyses than older legacy systems. Hadoop integrates cleanly with other enterprise data management systems. Moving data among existing data warehouses, newly available log or sensor feeds and Hadoop is easy. Hadoop is a powerful new tool that complements current infrastructure with new ways to store and manage data at scale.

MapReduce: Simplified Data Processing on Large Clusters

MapReduce is a programming model and software framework first developed by Google (Google's MapReduce paper submitted in 2004) intended to facilitate and simplify the processing of vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. Computational processing occurs on both:

- Unstructured data: file system.
- Structured data: database.

MapReduce framework

1. Per cluster node:
   1.1) Single JobTracker per master
      a. Responsible for scheduling the jobs' component tasks on the slaves.
      b. Monitors slave progress
      c. Re-executing failed tasks
   1.2) Single TaskTracker per slave
      a. Execute the tasks as directed by the master.

MapReduce Core Functionality:

1. Code usually written in Java- though it can be written inother languages with the Hadoop Streaming API.
2. Two fundamental pieces:
   a. Map step
      i. Master node takes large problem input and slices it into smaller sub problems; distributes these to worker nodes.
      ii. Worker node may do this again; leads to a multi-level tree structure
      iii. Worker processes smaller problem and hands back to master
   b. Reduce step
      i. Master node takes the answers to the sub problems and combines them in a predefined way to get the output/answer to original problem.
3. Data flow beyond the two key pieces (map and reduce):
   a. Input reader – divides input into appropriate size splits which get assigned to a Map function.
   b. Map function – maps file data to smaller, intermediate<key, value> pairs
   c. Partition function – finds the correct reducer: given the key and number of reducers, returns the desired Reduce node.
   d. Compare function – input for Reduce is pulled from the Map intermediate output and sorted according to this compare function.
   e. Reduce function – takes intermediate values and reduces to a smaller solution handed back to the framework.
   f. Output writer – writes file output.
4. A MapReduce Job controls the execution
   i. Splits the input dataset into independent chunks.
   ii. Processed by the map tasks in parallel.
5. The framework sorts the outputs of the maps.
6. A MapReduce Task is sent the output of the framework to reduce and combine.
7. Both the input and output of the job are stored in a file system.
8. Framework handles scheduling.

MapReduce Input and Output

1. MapReduce operates exclusively on <key, value> pairs.
2. Job Input : <key, value> pairs.
3. Job Output : <key, value> pairs. Conceivably of different types.
4. Key and value classes have to be serializable by the framework.
5. Default serialization requires keys and values to implement Writable.
6. Key classes must facilitate sorting by the framework.

Execution of Input and output parameters in typical MapReduce Framework

This execution of Map and Reduce algorithm is further explained in the implementation section.

Understanding Map and Reduce

Let us consider a simple problem wherein we have to search for a pattern 'cs396t' in a collection of files. We would typically run a command like this:
grep -r "cs395t" <directory>

Now, suppose you have to do this search over terabytes of data and you have a cluster of machines at your disposal? How can you make this grep faster?
Build a distributed grep!

Now the question arises, do we really need to consider a distributed grep? Why can't we just use our desktop for processing.Considering this in mind, let us estimate how much time will the average desktop system will take to process to search over terabytes of data.

In general, Considering an average read speed of 90MB/s: ~3.23 hours (Numbers are for Western Digital 1TB SATA/300 drive)

If you use an SSD with read speed of 350MB/s: ~50 minutes (Numbers are for Crucial 128 GB m4 2.5-Inch Solid State Drive SATA 6Gb/s)

This seems to be a huge amount of time considering the real time of data demanding requests from the internet. This is an approx. time only for searching through a collection of files. It would be huge amount of time when asked to sort a terabyte of data.

This definitely proves to be a wonder solution to the amount of time it takes to sort and work through huge collection of data. But, keeping this in mind, I could actually build a distributed system which does the same amount of work in this time. Can we ? The answer is an absolute NO!

Algorithmic Analysis :

```
var a = [1,2,3];
for (i=0; i<a.length; i++)
a[i] = a[i] * 2;
for (i=0; i<a.length; i++)
a[i] = a[i] + 2;
```

I can change it to:

```
function map(fn, a) {
for (i = 0; i<a.length; i++)
a[i] = fn(a[i]);
}
map(function(x){return x*2;}, a);
map(function(x){return x+2;}, a);
function sum(a) {
var s = 0;
for (i = 0; i<a.length; i++)
 s += a[i];
return s;
}
function join(a) {
var s = "";
for (i = 0; i<a.length; i++)
 s += a[i];
return s;
}
alert(sum([1,2,3]));
alert(join(["a","b","c"]));
function reduce(fn, a, init) {
var s = init;
for (i = 0; i<a.length; i++)
 s = fn( s, a[i] );
return s;
}
function sum(a) {
return reduce(function(a, b){return a+b;}, a, 0);
}
function join(a) {
return reduce(function(a, b){return a+b;}, a, “” );
}
alert(sum([1,2,3]));
alert(join(["a","b","c"]));
```

1. Passing functions as arguments – functional programming.
2. map – does something to every element in an array – can be done in any order! (amendable to parallelization)
3. So, if you have 2 CPUs, map will run twice as fast.
4. map is an example of embarrassingly parallel computation.

Suppose you have a huge array with elements which are all the webpages from the Internet. To search the whole internet:

1. you just need to pass a string_searcher function to map
2. reduce will be an identity function
3. run a MapReduce job on a cluster
4. that’s it! You are searching the Internet by writing just a few lines of code!

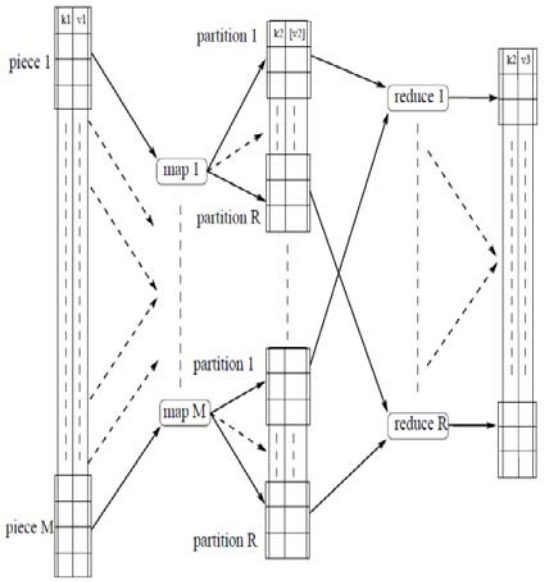
MAP- function that takes key/value pairs as input and generates an intermediate set of key/value pairs.

REDUCE- function that merges all the intermediate values associated with the same intermediate key.

User needs to define these two functions.

map: (k1, v1) □ list(k2, v2)

reduce: (k2, list(k2, v2)) □ list(v2)



EXAMPLE - WORD COUNT

Problem : counting occurrences of words in a large collection of documents.

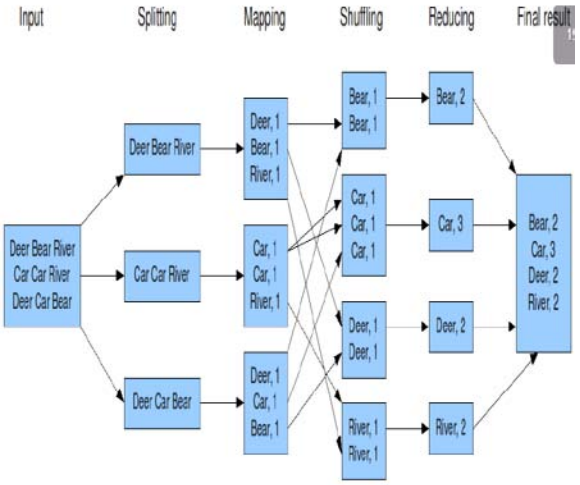
```
map(String key, String value):
 // key: document name
 // value: document contents
for each word w in value:
EmitIntermediate(w, "1");
reduce(String key, Iterator values):
 // key: a word
 // values: a list of counts
```

```
int result = 0;
for each v in values:
result += ParseInt(v);
Emit(AsString(result));
```

Other than map and reduce, user needs to provide:

1. names of input and output files
2. optional tuning parameters (size of split, M, R, etc.)

User's code is linked with MapReduce library and the binary is submitted to a task runner.



Word Counting using MapReduce

Other Examples of MapReduce :

1. Distributed grep
   a. map emits a line if it matches the given pattern
   b. reduce just copies input to output
2. Counting URL access frequency
   a. map processes web server logs and outputs <URL, 1>
   b. reduce sums all numbers for a single URL
3. Inverted index
   a. map function parses document and emits <word, docID>
   b. reduce gets all pairs for a given word and emits <word, list(docID)>

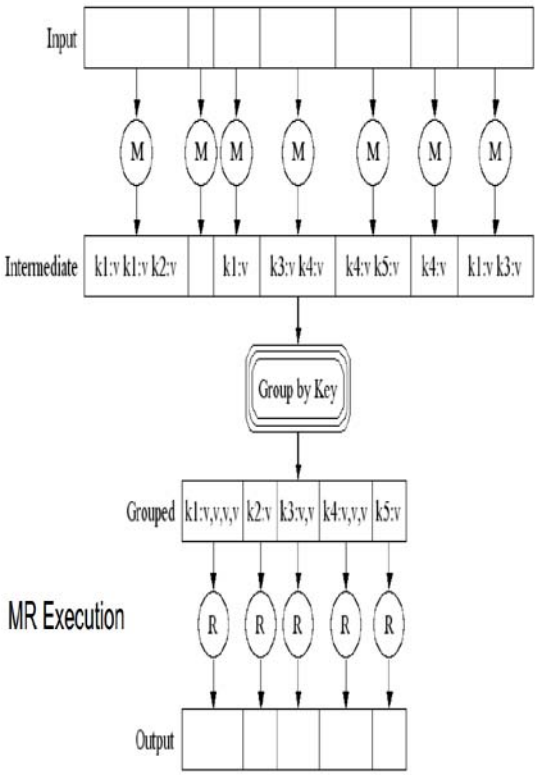
Implementing Map and Reduce in real world Scenarios :

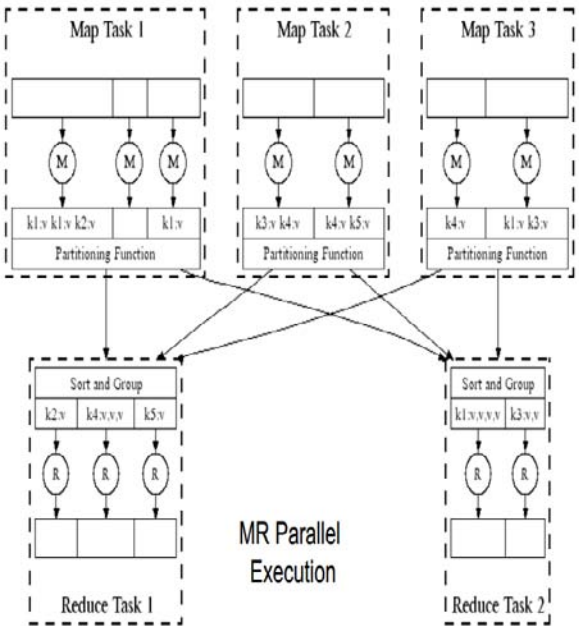
Map and Reduce can achieve the following achieving great scalability and speed.

1. Exploit parallelism in the computation.
2. Massively scalable – can run on hundreds or thousands of machines.
3. Hide the details of cluster management tasks like scheduling of tasks, partitioning of data, network communication from the user.
4. Fault tolerant (in large clusters failures are a norm rather than being an exception)

Opportunities for Parallelism using Map and Reduce:

1. Input – all key/value pairs can be read and processed in parallel by map
2. Intermediate grouping of data – essentially a sorting problem; can be done in parallel and results can be merged.
3. Output – All reducers can work in parallel. Each individual reduction can be parallelized.





MASTER : (in reference to fig 1)

1. Only 1 Master per MR computation.
2. Master:
   a. assigns map and reduce tasks to the idle workers.
   b. informs the location of input data to mappers.
   c. stores the state (idle, in-progress, completed) and identity of each worker machine.
   d. for each completed map task, master stores the location and sizes of intermediate files produced by the mapper; this information is pushed to workers which have in-progress reduce tasks.
3. Split the input into M pieces and start copies of program on different machines.
4. One invocation acts as the master which assigns work to idle machines.
5. Map task:
   a. read the input and parse the key/value pairs.
   b. pass each pair to user-defined Map function
   c. write intermediate key-value pairs to disk in R files partitioned by the partitioning function.
   d. pass location of intermediate files back to master.
6. Master notifies the reduce worker.
7. Reduction is distributed over R tasks which cover different parts of the intermediate key's domain.
8. Reduce task:
   a. read the intermediate key/value pairs.
   b. sort the data by intermediate key (external sort can be used)
   (note: many different keys can map to the same reduce task)
   c. iterate over sorted data and for each unique key, pass the key and set of values to user-defined Reduce function.
   d. output of Reduce is appended to final output for the reduce partition.
9. MR completes when all map and reduce tasks have finished.

MapReduce OUTPUT:

1. The output of MR is R output files (one per reduce task).
2. The partitioning function for intermediate keys can be defined by the user.

By default, it is "hash(key) mod R" to generate well balanced partitions.

3. Result files can be combined or fed to another MR job.

MapReduce Fault tolerance : Worker Failures

1. Master pings every worker periodically (alternatively, the worker can send a heartbeat message periodically)
2. If worker does not respond, master marks it as failed.
3. Map worker:
   a. any completed or in-progress tasks are reset to idle state.
   b. completed tasks need to be re-run since output is stored on a local file system
   c. all reduce workers notified of this failure (to prevent duplication of data)
4. Reduce worker:
   a. any in-progress tasks are reset to idle state.
   b. no need to re-run completed tasks since output stored in global file system.

Fault tolerance : Master Failure

1. Master periodically checkpoints its data structures.
2. On failure, new master can be elected using some leader election algorithm.
3. Theoretically, the new master can start off from this checkpoint.
4. Implementation: MR job is aborted if the master fails.

Fault tolerance : Network Failure

1. Smart replication of input data by underlying file system.
2. Workers unreachable due to network failures are marked as failed since its hard to distinguish this case from worker failure.
3. Network partitions can slow down the entire computation and may need a lot of work to be redone.

Fault tolerance : File System/Disk Failure

1. Depend on the filesystem replication for reliability.
2. Each data block is replicated f number of times. (Default : 3)

Fault tolerance: Malformed Input

1. Malformed input records could cause the map task to crash.
2. Usual course of action: fix the input.
3. But what if this happens at the end of a long-running computation?
4. Acceptable to skip some records (sometimes)
   a. Word count over very large data set.
5. MR library detects bad records which cause crashes deterministically.

Fault tolerance: Bugs in User Code

1. Bugs in user provided Map and Reduce functions could cause crashes on particular records.
2. This case similar to the failure due to malformed input.

Task Granularity:

1. M map tasks and R reduce tasks.
2. M and R much larger than the number of machines.
   a. Improves dynamic load balancing (add/remove machines)
   b. Speeds up recovery
      i. less work needs to be redone
      ii. I work already completed by a failed task can be distributed across multiple idle workers.

c. Bounds:
   1. Master makes O(M+R) scheduling decisions
   2. Master maintains O(M*R) state in memory.
   3. M is chosen such that each task works on one block of data.
   4. R is usually constrained by users to reduce the number of output files.

Requirements of applications using MapReduce

1. Specify the Job configuration
   a. Specify input/output locations
   b. Supply map and reduce functions via implementations of appropriate interfaces and/or abstract classes.
2. Job client then submits the job (jar/executablesetc) and the configuration to the JobTracker.

What are Hadoop/MapReduce limitations?

1. Cannot control the order in which the maps or reductions are run
2. For maximum parallelism, you need Maps and Reduces to not depend on data generated in the same MapReduce job (i.e. stateless)
3. A database with an index will always be faster than a MapReduce job on unindexed data.
4. Reduce operations do not take place until all Maps are complete (or have failed then been skipped)
5. General assumption that the output of Reduce is smaller than the input to Map; large data sourceused to generate smaller final values.

III. Conclusion

Traditional data processing and storage approaches are facing many challenges in meeting the continuously increasing computing demands of Big Data. This work focused on MapReduce, one of the key enabling approaches for meeting Big Data demands by means of highly parallel Processing on a large number of commodity nodes.

Issues and challenges MapReduce faces when dealing with Big Data are identified and categorized according to four main Big Data task types: data storage, analytics, online processing, and security and privacy. Moreover, efforts aimed at improving and extending MapReduce to address identified challenges are presented. By identifying MapReduce challenges in Big Data, this paper provides an overview of the field, facilitates better planning of Big Data projects and identifies opportunities for future research.

References RéférencesReferencias

1. H. Yang and S. Fong, "Countering the concept-drift problem in Big Data using iOVFDT," IEEE International Congress on Big Data, 2013.
2. S. Ghemawat, H. Gobioff and S. Leung, "The Google file system," ACM SIGOPS Operating Systems Review, 2003.
3. J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, 51(1), pp. 107-113, 2008.
4. Apache Hadoop, http://hadoop.apache.org
5. Z. Xiao and Y. Xiao, "Achieving accountable MapReduce in cloud computing," Future Generation Computer Systems, 30,pp. 1-13, 2014.
6. W. Zeng, Y. Yang and B. Luo, "Access control for Big Data using data content," IEEE International Conference on Big Data, 2013.
7. C. Parker, "Unexpected challenges in large scale machine learning," Proc. of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, 2012.
8. www.ibm.com/software/data/infosphere/hadoop/mapreduce/
9. hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
10. research.google.com/archive/mapreduce-osdi04.pdf

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 9 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Quality of Service Centric Web Service Composition: Assessing Composition Impact Scale towards Fault Proneness

By Sujatha Varadi & Dr. G. Appa Rao

GITAM University, India

Abstract- Service composition in service oriented architecture is an important activity. In regard to achieve the quality of service and secured activities from the web service compositions, they need to be verified about their impact towards fault proneness before deploying that service composition. Henceforth, here in this paper, we devised a novel statistical approach to assess the service composition impact scale towards fault proneness. The devised model explores the higher and lower ranges of the service composition impact scale, which is from the knowledge of earlier compositions that are notified as fault prone.

Keywords : *web service compositions, composition support, service composition impact scale, service descriptor impact scale, web service composition fault proneness.*

GJCST-C Classification : *H.3.5*

Strictly as per the compliance and regulations of:

Quality of Service Centric Web Service Composition: Assessing Composition Impact Scale towards Fault Proneness

Sujatha Varadi [α] & Dr. G. Appa Rao[σ]

Abstract- Service composition in service oriented architecture is an important activity. In regard to achieve the quality of service and secured activities from the web service compositions, they need to be verified about their impact towards fault proneness before deploying that service composition. Henceforth, here in this paper, we devised a novel statistical approach to assess the service composition impact scale towards fault proneness. The devised model explores the higher and lower ranges of the service composition impact scale, which is from the knowledge of earlier compositions that are notified as fault prone. The experimental results explored from the empirical study indicating that the devised model is significant towards estimating the fault proneness scope of any service composition from selected service descriptors.

Keywords: web service compositions, composition support, service composition impact scale, service descriptor impact scale, web service composition fault proneness.

I. Introduction

Service-Oriented Architecture (SOA) simplifies information technology related operational tasks by consumption of ready-to-use services. Such SOA found to be realized currently in ecommerce domains such as B2B, B2C, C2B and C2C, in particular the web services are one that considered serving under this SOA.

Web services are software components with native functionality that can be operable through web. Another important factor about this web services is that more than one service can be composed as one component by coupled together loosely. The standard WSDL is web service descriptive language that let the self exploration of the web services towards their functionality and UDDI is the registry that lets the devised web services to register and available to required functionality [1].

Composition of web services is loosely interconnected set of Web service operations that acts as a single component, which offers solutions for divergent tasks of an operation.

Author α: Assistant Professor, Department of CSE GITAM Schoolof Technology, GITAM University, Hyderabad.
e-mail: varadi.sujatha@gmail.com
Author σ: Professor, Department of CSE GITAM Institute of Technology, Visakhapatnam. e-mail: apparao_999@yahoo.com

Since the task of composition is integrating divergent web services explored through different descriptors, it is the most fault prone activity. The functionality of service composition includes the activities such as (i) identify the tasks involved in a given business operation, (ii) trace related web services to fulfill the need of each task, (iii) couple these services by exploring the order of that services usage, which is based on the expected information flow, (iv) and resolve the given operation by ordering the responses of the web services that coupled loosely as one component.

In order to achieve quality of service and secure transactions in web service composition and usage, the impact of the composition should be estimated before deploying those loosely coupled web services as one component.

The Web service compositions used earlier that can be found in repositories and the services involved in those compositions helps to assess the impact of these web services towards fault proneness.

The current composition strategies [2] [3] [4] [5] [6] [7] [8] are error prone, since these State-of-the-art techniques are not mature enough to guarantee the fault free operations. However, finding these compositions as fault prone after deployment is functionally very expensive and not significant towards end level solutions, also may leads to serious vulnerable. Hence the process of estimating the composition scope towards fault proneness is mandatory.

In this paper, we propose a novel statistical approach to estimate the impact scale of a service composition towards fault proneness. Our approach acts as an assessment strategy for any of existing web service composition approaches.

The paper is structured as follows. Section 2 discusses related work. In section 3, the proposed statistical approach is explored, which followed by Section 4 that contains the results explored from empirical study. The conclusion of the proposal and future research directions were discussed in Section 5.

II. Related Work

Service compositions with malfunctioned web services lead to form the highly fault prone

compositions. Henceforth the web service composition to serve as one component under SOA is complex and needs research domain attention to deliver effective strategies towards the QoS centric service. The model devised in [9] defined set of QoS factors to predict feasible services. Many of existing quality-aware service selection strategies aimed to select best service among multiple services available. The model devised in [8] considering the linear programming to find the linear combination of availability, successful execution rate, response time, execution cost and reputation, which is in regard to find the optimal service composition towards given business operation. The model devised in. [6] is considering the temporal validity of the service factors. The authors in [10] modeled a mixed integer linear program that considers both local and global constraints.

The model devised in [7] is selecting services as a complex multi-choice multi-dimension rucksack problem that tends to define different quality levels to the services, which further taken into account towards service selection. All these solutions are depends strongly on the positive scores given by users to each parameter. However, it is not scalable to establish them in prospective order.

Though the QoS strategies defined are used in service composition the factor fault proneness of the service composition is usual. In regard to this a model devised in [11] explored a mechanism for fault proliferation and resurgence in dynamically connected service compositions. Dynamically coupled architecture outcomes in further complexness in need of fault proliferation between service groups of a composition accomplished by not depending on other service groups.

In a gist, it can be conclude that almost all of the benchmarking service quality assessment models are attribute specific, user rating specific or both. Hence importance of attributes is divergent from one composition requirement to other, and the user ratings are influenced by contextual factors, and another important factor is all of these bench mark models are assessing services based on their individual performance, but in practice the functionality of one service may influenced by the performance of other service. Henceforth here in this paper we devised a statistical approach that estimates the impact scale of service composition towards fault proneness, which is based on a devised metric called composition support of service compositions and service descriptors.

III. Estimating the Service Composition Impact Scale Towords Fault Proneness

The said statistical model works in two aspects. First, it estimates the impact of each web service descriptor to form a selected malfunctioned service composition. And then it estimates the higher and lower ranges of the impact scale o towards fault proneness, which is from the impact of each service descriptor and each malfunctioned service composition. Then these higher and lower ranges of the impact scale will be used to assess the impact of a newly composed service composition towards fault proneness. This strategy leads to estimate the problem of web service descriptor selection. The business solution expected might represented by several compositions, but selecting one of these compositions is strictly by their impact towards fault proneness. The proposed model is optimal in this regard. The detailed exploration of the proposed model is as follow:

The approach of measuring Composition support ( ) metric is proposed in this paper. In regard to measure the composition support, we consider the bipartite graph that represents the composition weights.

a) Assumptions

Let set of service-composites $wsc_1, wsc_2, wsc_3, ...wsc_n$, which found to be malfunctioned compositions

Let set of web service descriptors $wsd_1, wsd_2, wsd_3, ..., wsd_n$, which were involved to form compositions opted

Hereafter the set of such web-service descriptor sets will be referred as

Let two web-service descriptors wsd_i and wsd_j, wsd_i connected with wsd_j, if and only if $(wsd_i, wsd_j) \in wsc_i$

Build an undirected weighted graph UWG with web-service descriptors as vertices and edges between web-services descriptors. An edge between the two web-service descriptors will be weighted as follows

$foreach\{wsds \forall wsds \in SWSS\}$

$$ew_{(wsd_i \leftrightarrow wsd_j)} = \frac{\sum_{k=1}^{|SWSS|} \{1 \exists [(wsd_i, wsd_j) \subseteq wsc_k \wedge i \neq j]\}}{|SWSS|}$$

Here in the above equation $ew_{(wsd_i \leftrightarrow wsd_j)}$ indicates the edge weight between web-service descriptors wsd_i and wsd_j

In the process of building a weighted graph we consider that an edge between any two web-service descriptors exists if and only if the edge weight $ew > 0$

b) Process

In the process of detecting the composition support of each web-service descriptor with service-compositions, initially we build a bi-parted graph between web service compositions and the set of web-service descriptors.

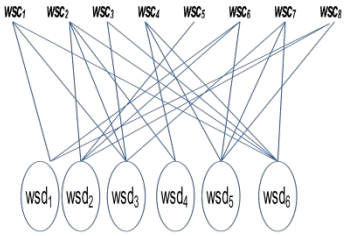


Figure 1: bipartite graph between web service compositions and web-service descriptors

If a web-service descriptor wsd_i is part of a web-service composition wsc_i then the weight of the connection between wsd_i and wsc_i will be measured as follows:

$$cw_{(wsd_i \leftrightarrow wsc_j)} = \frac{\sum_{k=1}^{|wsc_i|} \{ew_{(wsd_i \leftrightarrow wsd_k)} \exists [i \neq k \wedge (wsd_i, wsd_k) \in wsc_j\}}{|wsc_j|}$$

Here in the above equation we consider the sum of all edge weights from undirected graph such that there exists an edge between web service descriptor wsd_i and other descriptors of the web service composition wsc_j. The '$|wsc_j|$' indicates the total number of descriptors in web service composition wsc_j.

The graph representation (fig. 1) indicates the bipartite relation between web-service descriptors and web service compositions. Composition weights of the different web service compositions represent their importance. Intuitively, a web service composition with high composition weight should contain many of the web-service descriptors with high composition support. The underpinning association of web service compositions and web-service descriptors is that of association between hubs and authorities in the HITS model [13].

The devised process of identifying web service composition weights using bipartite graph is explored below:

Let consider a matrix format of the connection weights of the bipartite edges between web-service descriptors and web-service compositions in given bipartite graph.

The weight of the each web service composition as a hub in a bipartite graph is initialized as 1, which we represented as matrix (table 1).

Table 1: Initializing the weight of the each web service composition as hub in bipartite graph with 1 and represented them as a matrix u as follows.

1
1
1
1
1
1

Let the weights between descriptors and compositions of the given bipartite graph (see fig 1) and form a matrix such that rows represent descriptors (authorities) and columns represent compositions (hubs) and refer that matrix as A,

As referred in HITS [13] algorithm, find each web service descriptor (authority) weight, which is can be done as follows:

$$v = A^{'} Xu$$

Here in the above equation v is the matrix representation of the web service descriptor weights as authorities, $A^{'}$ is the transpose matrix of the matrix A, which is the matrix representation of connection weights between web service compositions as hubs and web service descriptors as authorities in bipartite graph. Then the actual weights of the web service compositions (hubs) can be measured as follows:

$$u = AXv$$

The matrix multiplication between matrix A and matrix v results the actual weights of the service compositions as hubs.

Then the composition support cs of web-service descriptor wsd can be measured as follows

$$cs_{wsd} = \frac{\sum_{i=1}^{m} \{u_{wsc_i} \exists cw_{wsd \leftrightarrow wsc_i} > 0\}}{\sum_{i=1}^{m} u_{wsc_i}}$$

And then web service composition impact scale towards fault proneness of each service-composition can be found as follows:

$$\sigma_{wsc_i} = 1 - \frac{\sum_{j=1}^{m} \{cs_{wsd_j} \exists wsd_j \in wsc_i\}}{|WSD|}$$

Here in the above equation $|WSD|$ indicates the total number of web-service descriptors involved to create all web service compositions.

Then the web service composition impact scale threshold τ towards fault proneness can be measured as follows:

$$\tau = \frac{\sum_{i=1}^{|SWSS|} \sigma_{wsc_i}}{|SWSS|}$$

Here in the above equation $|SWSS|$ indicates the total number of service-compositions considered

Then the standard deviations of the σ each service composition from τ will be measured further, which is as follows:

$$sdv_{\tau} = \sqrt{\frac{\left(\sum_{i=1}^{|SWSS|} \left(\sigma_{wsc_i} - \tau\right)^2\right)}{(|SWSS| - 1)}}$$

Then the Web service composition impact scale low and high ranges towards fault proneness are explored as follows

Lower range of impact scale τ_l is

$$\tau_l = \tau - sdv_{\tau}$$

Higher range of impact scale τ_h towards fault proneness is

$$\tau_h = \tau + sdv_{\tau}$$

Service-composite can be said as safe if and only if $\sigma_{wss} < \tau_l$

The impact scale of service composition wsc towards fault proneness is high if and only if

$$\sigma_{wsc} \geq \tau_l \,\&\&\, \sigma_{wss} < \tau_h$$

The service composition is said to be fault prone if

$$\sigma_{wsc} > \tau_h$$

IV. Empirical Analysisand of the Proposed Model

This work explored the credibility of the proposed model on set of 296 service compositions.

The above said data set contains 294 samples, out of that 250 samples were used to devise the Degree of fault prone threshold and its upper and lower bounds. Further we used the rest 44 records to predict the fault proneness scope. Interestingly, the empirical study delivered promising results. The statistics explored in table 10

Table 10: Statistics of the experiment results

Total Number of web service composites	296
Total number web service descriptors used	140
Total number of edges determined	1560
Total number of bipartite edges found	27776
Service composition impact scale threshold τ towards fault proneness	0.46795646260519363
Higher range of τ	0.5284095974190264
Lower range of τ	0.4075033277913609

Table 2: Exploration of the parameters used in empirical study

Among the considered web service compositions, 244 web service compositions were used to estimate the service composition impact scale towards fault proneness

Total web service composites used to test the accuracy of the impact scale are 56

Total number of false negatives are 11, that is web service composites found with σ less than lower bound are 11

Total number of true positives found is 41, which are having σ greater than lower bound.

a) Performance Analysis

We used accuracy estimation (the percentage of valid predictions by the proposed) as the main performance measure. In addition to measuring accuracy, the precision, recall, and F-measure were used to analyze the performance; these are defined using following equations.

$$pr = \frac{t_+}{t_+ + f_+}$$

Here in above Equation the pr indicates the precision, t_+indicates the true positives and f_+indicates the false positive

As per the empirical study conducted the t_+found here are 41 and $f_+$are 0, henceforth precision is 1.

$$rc = \frac{t_+}{t_+ + f_-}$$

Here in above Equation, the 'rc' indicates the recall, f_-indicates the false negative. As per the results explored in empirical study $f_-$are 11, hence the rc value is 0.788.

$$F = \frac{2 * pr * rc}{pr + rc}$$

Here in the above Equation, Findicates the F-measure. And the F-measure found from the results of the empirical study is 0.88143

As per the results explored, the proposed model is accurate to the level of 79%. The failure percentage is 21%, which is not negligible but considerably performed well.

V. Conclusion

The model devised in this paper is a method of estimating web service composition impact scale

towards fault proneness. This approach is a statistical analysis that derives lower and higher range of service composition impact scale towards fault proneness. In regard to this initially an undirected graph that connects the involved web service descriptors as vertices with weighted edges. The edge weight of to vertices is the ratio of service compositions contains services from both descriptors act as vertices to a selected edge. Further a bipartite graph build between web service compositions as hubs and web service descriptors used to compose those compositions as authorities. Further hub and authority weights were calculated as explored in section 3, and further these weights were used to estimate the service composition impact scale towards fault proneness. The estimated service composition impact scale higher and lower range values can be used further to estimate the impact of any service composition towards fault proneness. The empirical analysis was conducted on dataset with 296 divergent web service compositions. The explored results are indicating the significance of the proposed model. In future to improve the accuracy of the devised model, the correlation of the service descriptors will be estimated, which is done by considering the web-services of each descriptor as categorical value set. Further, web-service reputation can also be considered to estimate the impact of a service composition towards fault proneness.

References Références Referencias

1. Papazoglou, M. P., Georgakopoulos, D.: "Service-oriented computing", Communications of the ACM, Vol. 46, No. 10, 2003, pp. 25–28.
2. Aggarwal, R., et al.: "Constraint-driven Web Service Composition in METEOR-S", IEEE Conference on Service Computing, 2004.
3. Lazovik, A., Aiello, M., Papazoglou, M.: "Planning and monitoring the execution of web service requests", International Conference on Service-Oriented Computing, 2003, pp. 335-350.
4. Sirin, E., Hendler, J., Parsia, B.: "Semi-automatic Composition of Web Services Using Semantic Descriptions", In Web Services: Modeling, Architecture and Infrastructure workshop in ICEIS, 2003.
5. Srivastava, B., Koehler, J., "Web Service Composition - Current Solutions and Open Problems", Proceedings of ICAPS Workshop on Planning for Web Services, 2003.
6. Martin-Diaz, O., Ruize-Cortes, A., Duran, A., Muller, C.: "An Approach to Temporal-Aware Procurement of Web Services", International Conference on Service-Oriented Computing, 2005, pp. 170–184.
7. Yu, T., Lin, K.J.: "Service Selection Algorithms for Composing Complex Services with Multiple QoSConstraints", International Conference on Service-Oriented Computing, 2005, pp. 130–143.
8. Zeng, L., Benatallah, B., et al.: "QoS-aware Middleware for Web Services Composition", IEEE Transactions on Software Engineering, Vol. 30, No. 5, 2004, pp. 311–327.
9. Cardoso, J., Sheth, A., Miller, J., Arnold, J., Kochut, K.: "Quality of service for workflows and web service processes", Journal of Web Semantics, Vol. 1, No. 3, 2004, pp. 281–308.
10. Ardagna, D., Pernici, B.: "Global and Local QoS Constraints Guarantee in Web Service Selection," IEEE International Conference on Web Services, 2005, pp. 805–806.
11. Chafl, G., Chandra, S., Kankar, P., Mann, V.: "Handling Faults in Decentralized Orchestration of Composite Web Services", International Conference on Service-Oriented Computing, 2005, pp. 410–423.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 9 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Nomenclature and Benchmarking Models of Text Classification Models: Contemporary Affirmation of the Recent Literature

By Venkata Ramana. A & Dr. E. Kesavulu Reddy
S.V.University, India

Abstract- In this paper we present automated text classification in text mining that is gaining greater relevance in various fields every day. Text mining primarily focuses on developing text classification systems able to automatically classify huge volume of documents, comprising of unstructured and semi structured data. The process of retrieval, classification and summarization simplifies extract of information by the user. The finding of the ideal text classifier, feature generator and distinct dominant technique of feature selection leading all other previous research has received attention from researchers of diverse areas as information retrieval, machine learning and the theory of algorithms. To automatically classify and discover patterns from the different types of the documents [1], techniques like Machine Learning, Natural Language Processing (NLP) and Data Mining are applied together. In this paper we review some effective feature selection researches and show the results in a table form.

GJCST-C Classification : *H.1, E.4, H.2.1*

Strictly as per the compliance and regulations of:

Nomenclature and Benchmarking Models of Text Classification Models: Contemporary Affirmation of the Recent Literature

Venkata Ramana. A [α] & Dr. E. Kesavulu Reddy [σ]

Abstract- In this paper we present automated text classification in text mining that is gaining greater relevance in various fields every day. Text mining primarily focuses on developing text classification systems able to automatically classify huge volume of documents, comprising of unstructured and semi structured data. The process of retrieval, classification and summarization simplifies extract of information by the user. The finding of the ideal text classifier, feature generator and distinct dominant technique of feature selection leading all other previous research has received attention from researchers of diverse areas as information retrieval, machine learning and the theory of algorithms. To automatically classify and discover patterns from the different types of the documents [1], techniques like Machine Learning, Natural Language Processing (NLP) and Data Mining are applied together. *In this paper we review some effective feature selection researches and show the results in a table form.*

I. INTRODUCTION

Research on text categorization has emerged into a new level with the speed in advancement of internet technology. Various techniques were developed such as Machine Learning, Support Vector Machines (SVMs), KNN, Neural Network, Boosting and Naive Bayes variants [2] etc. Machine learning technique has evolved into a foremost model of text categorization [3].

Text classification is used in diverse fields for managing documents stored as texts in databases. The information today is composed of a large set of documents from multiple sources, such as news, articles, books, digital libraries, e-mail messages and web pages. Applications of text classification are being used in areas such as; in news delivery for classifying articles automatically into subjects and made available to users based on their search profile or interests.

In content management, grouping documents into many-sided categories is to simplify searching and browsing. In identifying spam mail where thequestionable mails are flagged as suspected spam and separated for batch deletion. In e-commerce for item descriptions in shopping and auction web sites where short texts are used for classification. In call centers offering support services, the text notes of call logs are classified with respect to defined criteria to identify trends periodically. These are but a few examples of how text classification is finding its way into applications and text classification systems.

The text classification process involves various steps like indexing, feature generation, feature filtering and feature selection.

a) Document Preprocessing

Document preprocessing step indexes the documents to minimize the complexity in documents.

b) Feature Selection

Feature selection methods are a preprocessing step. The selected features from the training set are then used to classify new incoming documents. The popular feature selection methods are document frequency, term frequency, chi-square statistic and Accuracy. Feature selection consists of the following steps;

i. Preprocessing - In the preprocessing stage, the various steps are;
   a. Feature Extraction - To generate text features based on the occurrence of the words in the document.
   b. Feature Filtering - To eliminate irrelevant or noisy features and non-discrimination in the data [4] effectively reducing the size of feature set.
   c. Document Representation - The document is transformed from a full text version to a vector space representation.

ii. Classifier building - Is performed by a score based strategy where words are scored with respect to their occurrence in the given document. This is predefined by the measure of weight of the word to decrease the high dimensionality space.

The selected features are then used to apply feature selection using different policies for text classification.

In this paper we discuss feature selection methodologies in text classification and review some effective methods for text classification and how performance can be increased [5, 6].

Author α: Research Scholar, Department of Computer Science S.V.University,Tirupat-India-517502. e-mail: avr_rdg@yahoo.co.in
Author σ: Assistant Professor, Department of Computer Science S.V.University,Tirupat-India-517502.
e-mail: ekreddysvu2008@gmail.com

II. Nomenclature of the Feature Selection Strategies for ext Classification

a) Text Classification

Text classification is the process of classifying the documents into predetermined categories. The engineering methodologies define a group of consistent rules for accurately classifying documents into a categorical set. The nature of classifying the documents can be of 3 types: i) Unsupervised, ii) Supervised and iii) Semi Supervised. Automatic text classification widely studied since the last few years has rapidly evolved due to fast development of internet technology.

1. Document Preprocessing - Document preprocessing step indexing of documents to minimize the complexity in the documents.
   The documents are classified into 2 types based on the class:
   (i) Single Label and
   (ii) Multi-Label.
   Single label document are fit for only a single class whereas multi label documents may be fit for single and multiple classes.
2. Feature Selections - Feature selection involves generation of filters focusing on relevant and informative data. The feature filtering process and feature selection are used for selecting features useful for text categorization with respect to scalability, efficiency as well as accuracy.
3. Applying Feature Selection: To apply feature selection in text categorization there are two major policies;
   (i) Local Policy: In the local policy a varied set of features are chosen from each class not dependent on other classes providing identical weight to each class and optimizing the performance by choosing the most important features for each class.
   (ii) Global policy: In the global policy a single set of features are chosen from all classes, providing a global view of the whole dataset and a single global score from the local scores [8, 9].
   E.g.: The automatic labeling of every inbound news item with a predefined subject like "media", "politics" or "sports" or "art". First a training set D = (d1….. dn) of documents that are already labeled with a class C1, C2 (e.g. politics, sport) is selected. Next a classification model that is competent enough for labeling a new document d of the vertical with the right class is determined. The most commonly used document representation is known as vector space model (SMART) [7]. *In this paper we focus on classification of single label document.*
4. The challenges related with text classification come from many fronts and are mostly of three types;
   (i) Selection of a suitable data structure to represent the documents.
   (ii) Selection of right objective functions to prevent high formal dimensionality of the data and consequent algorithmic issues. It mainly leads to over-fitting where a classifier fits the training dataset well but performs inadequately on cases exterior to the training dataset that result in high computational overheads and increased training period.
   (iii) Text classification problems at times have only very limited training data present that poses a high difficulty for learning and the classification.

b) Text Feature Generators and Feature Extraction

The process of Text Feature Generation is to come up with an array of feature generators that make the feature selector choose powerful predictive features.

First it is important to ascertain what qualifies to be counted as a word or a term. Difficulty arises in instances like 'HP-UX' qualifying as single word or couple of words and how to ascertain the type of term '650-857-1501'? In programming, a simple solution requires contiguous sequence of alphabetic characters; or alphanumeric characters including identifiers like 'ioctl32', that are occasionally helpful. By means of the Posix regular expression \p{L&}+ we evade breaking 'naıve' in two and also several accented words in French, German, etc. Difficulty arises in case of words like 'win 32', 'can't' or words that may be hyphenated over a line break. Similar to several other data cleaning techniques, the list of exceptions is never-ending and we have to limit the expectation and expect for an 80%-20% exchange. An advantage is semantic errors in word parsing are generally observed by the core learning algorithm. So their statistical properties are of importance not its readability or intuitiveness to people. Major feature generators that are useful differ according to the domain text qualities. The typical feature generators applied are;

i. *Word Merging*

Merging is a technique of decreasing the size of the feature space considerably by merging different word variants and treating the new entity as a single feature. This significantly improves the predictive value of certain features.

Force conversion of all letters to lowercase is a generally accepted technique. Letters at the start of a sentence, which does alter the word's meaning, and helps lessen the dispersion. In case of proper nouns, it frequently conflates other word meanings, e.g. 'Bush' or 'LaTeX.'

Several word stemming algorithms could be used for merging multiple related word forms. For

instance, 'cat,' 'cats,' 'catlike' and 'catty' can all be merged into a common feature. Stemming normally is advantageous for recall however affects the precision. As in the example above if one is searching for 'catty' and the word is considered the same as 'cat,' then essentially a definite amount of precision is lost. In case of exceedingly skewed class distributions, this loss may greatly affect precision.

Also stemming algorithms give errors of both over-stemming and under-stemming; however, the semantics occupy less significance compared to feature's statistical properties. Stemmers have to be individually designed for every natural language and there are several fine stemmers existing for Roman languages whereas for other languages such as Hebrew and Arabic the stemming process becomes difficult. Further for some applications of text classification, there occurs mixing of multiple natural languages together, at times even inside only one training case. This necessitates a language recognizer to determine the type of stemming algorithm that needs to be used for each case or each sentence. However this level of complexity and slowdown is not desired. Stemming by merely considering the initial few characters of every word may give equal classification accuracy for several classification problems. Misspellings that commonly occur in technical texts or blogs and resolving with an automatic spelling correction step provided in the processing pipeline is occasionally proposed to ease classification but the errors revealed may overshadow the supposed advantage. A familiar problem with the spell checker is that out-of-vocabulary (OOV) words are forced to the nearest known word that can have totally different meaning. This is generally seen with technical terms that can be crucial predictors. Common misspellings may give frequent misspelled form and appear as a useful feature, e.g. 'volcano.'

Out-of-vocabulary (OOV) words are mostly words like abbreviations and acronyms found in governmental or technical texts. In case if glossaries can be referred, the short and long forms can be merged into a single term. Though several acronym dictionaries exist online, there are many types of short acronyms that are extremely document and even domain-specific. A few of the researches have shown success identifying acronym definitions in text, such as '(OOV)' above, that gives a locally clear-cut definition for the term. Online thesauruses may as well be used to merge together dissimilar words, e.g. to resolve the 'color' vs. 'hue' problem however the technique hardly ever helps, as multiple meanings exist for many words that distorted their final meanings. To disambiguate word meanings correctly would require a much deeper understanding of the text than is needed for text classification. However, this problem is overcome by using domain-specific thesauruses of synonyms. For instance in representing a huge set of part numbers corresponding to a common product line a single feature to represent all proves very beneficial.

While merging related words with each other can turn out features with additional frequent occurrence (characteristically with greater recall and lower precision).

ii. *Word Phrases*

Identifying multiple word phrases as a single term can generate rarer, highly specific features (which regularly aid precision and have lower recall), e.g. 'John Denver' or 'user interface.' However instead of using a dictionary of phrases as in the above case an easy technique is to consider all successive pairs of words as a phrase term and let feature selection decide which are helpful for prediction. Modeled on the new technology of online searching, the recent trend to eliminate spaces in proper names, e.g. 'SourceForge,' gives the specificity of phrases devoid of any particular software deliberations. Also the same can be applied to phrases having three or more words with intermittently more specificity and also with strictly decreasing frequency. Maximum advantage is gained with two-word phrases [10] to some extent as the parts of the phrase may previously have the identical statistical properties, e.g. the phrase with four words 'United States of America' is previously enclosed by the two-word phrase 'United States.' Also, the extent of a two-word phrase can be extended by removing general stopwords, e.g. 'head of the household' turns into 'head household.' However the stop word lists are language specific and have limitations. The main advantage of classification lies in increasing the extent of phrases, instead of removing frequently useless words that could be already removed in a language-independent fashion with maximum feature selection techniques.

iii. *Character N-grams*

The word identification techniques discussed previously do not succeed in some cases and cannot succeed in spotting some good features. For instance, languages like Chinese and Japanese do not employ a space character. Segmenting such text into words is difficult, however approximately comparable accuracy might be gained just by means of every set of adjoining Unicode characters as features $n-grams$. Definitely several of the variants will be worthless; however feature selection is able to identify the maximum predictive features. In case of languages that utilize the Latin character set, 3-grams or 6-grams may be right. For example, $n-grams$ would obtain the real meaning of common technical text patterns such a $HP-UX$ 11.0','while(◁▷){',#!/bin/',and':)' .Phrases of two adjoining $n-grams$ simply equate to $(2n)$ grams. The number of potential increase exponentially with where in reality it is merely a little fraction of the

possibilities that arise in actual training examples and only a small part of those could be predictive.

The general Adobe PDF document format, records the position of each character on the page and does not clearly records spaces. Software libraries to pull out the text from PDF utilize heuristics to determine where to output a space character. Due to this reason text extracts either occasionally overlook spaces amid the words or have a space character placed among every pair of letters. Obviously, such issues will cause chaos with a classifier that relies on spaces to recognize words. A more strong technique is for the feature generator to remove all whitespace that gives n-grams from the resulting sequence.

iv. *Multi-Field Records*

Multi-field records are regularly used as maximum applications have multiple text (and non-text) fields with respect to each record, though in most applications of text classification research deal with training cases as a single string. These fields in document management are usually title, author, abstract, key-words, body and references. In the field of technical support, these may be title, product, keywords, engineer, customer, symptoms, problem description, and solution. Additionally classifying long strings, e.g. arbitrary file contents, the first few kilobytes are usually taken as a separate field and that is generally adequate for generating desired features without the need to deal with huge files like star or zip archives.

The simplest approach is to concatenate all strings together. However, supposing the classification goal is to separate technical support cases by product type and model, then the most informative features may be generated from the product description field alone, and concatenating all fields will tend to water down the specificity of the features.

Another uncomplicated strategy is to provide every field with its own separate bag-of-words feature space. For example the word 'OfficeJet' given as the title field would be addressed as if it were not connected to a feature for the similar word in the product field. Occasionally multiple fields are required to be combined, and at the same time set aside as separate and while the rest are isolated. Such choices are done manually and an automated search improves computation time for the search and essentially reduces the expert's time, and it may identify better options not possible in manual search.

v. *Other properties*

In case of certain classification issues, text properties other than words or $n-grams$ generate the key predictors for high accuracy. Certain kinds of spam use deceptions such as '4*ree* v!@gr@4 u!' 'to prevent word-based features though these might easily be found by features identifying their abnormal word lengths and the density symbols. Similarly identifying Perl or awkcode, is done with specific alphanumeric identifiers, less specific in occurrence than the distribution of particular keywords and special characters. Information of formatting like the amount of whitespace, the word count, or the average number of words per line can be key features for specific tasks.

Here task-specific features created are usually extremely expensive like parsing particular XML structures that hold name-value pairs. The features being task-specific, it is especially difficult to provide common useful comments about their generation or selection. The insufficient information available regarding task-specific features in literature of text classification overrides their true importance in many practical applications.

vi. *Feature values*

Next with the determination of the word that could be considered as a feature term, the significance of the numerical feature must be ascertained. For certain purposes a binary value is enough to represent if the term actually occurs. This depiction is employed by the Bernoulli formulation of the Naive Bayes classifier [11]. A lot of other classifiers utilize the term frequency $tf_{t,k}$ (the word count in document k) directly as the feature value, e.g. the Multinomial Naive Bayes classifier [11].

The support vector machine (SVM) has demonstrated to be extremely effective in text classification. In such kernel techniques, the distance between two feature vectors is normally calculated as their dot product (cosine similarity), which is dominated by the dimensions with larger values. In order to avoid a situation where the extremely frequent but non-discriminative words (such as stop-words) dominate the distance function, we can either use binary features or weight of the term frequency value $tf_{t,k}$ inversely to the feature's document frequency df_t in the corpus (the number of documents in which the word appears one or more times). In this way, very common words are downplayed. This technique commonly known as TF.IDF has a many variants, one form being $tf_{t,k} \times \log\left(\frac{M+1}{df_t+1}\right)$, where M is the number of documents. Though this technique necessitates greater computation and storage per feature compared to binary features, it may further offer superior accuracy for kernel methods. The document length typically varies according to the document type short or long word counts with the same topic. To make these feature vectors more comparable, the $tf_{t,k}$ values can be normalized to make the length (Euclidean norm) of every feature vector equals to 1.

c) Feature Filtering

Feature selection process scores each possible feature with respect to a specific feature selection metric and selecting the most excellent k features. As discussed in the previous sections, a wide array of feature generation approaches, we now concentrate on feature filtering.

Selection of the best feature differs extensively from job to job where a number of values ought to be used. Keywords are extracted and examined according to criteria such as the incidence of repeated words or frequently used terms not definite to any category to eliminate irrelevant and noisy information. Feature filtering with respect to the training class labels scores each feature independently. The scoring counts the number of feature examples in training positive- and negative-class separately and next over these it computes a function. Increase in the number of features results in increase in the time necessary for initiation and training time affecting the accurateness of the classifier. To achieve dimensionality reduction the two most common approaches in machine learning or data mining are the filter and the wrapper [8, 12].

1. The filter chooses a subset of features by filtering based on scores assigned by specific weighting and is independent of any learning algorithm.

i. First in the process, the a) rare words and b) common words are removed;

a. The rare words can be eliminated, since they may not have any presence in future classifications. For instance, words with presence less than two times can be eliminated. Word frequencies characteristically pursue a *Zipf* distribution: the frequency of each word's incidence is proportional to where rank is its rank among words sorted by frequency, and is a fitting factor close to 1.0 (Miller 1958) [13]. Since of the total distinct words, a part equal to half of the total number can appear only once, so deleting terms below a specific low rate of incidence generates great savings. The meticulous choice of threshold value may affect the accuracy. If we remove rare words with respect to the count of the entire dataset prior to splitting off a training set, it will result in leaking a part of the information regarding the test set to the training phase. Avoiding major resource allocation for cross-validation studies, the research creates feasibility since avoiding class labels of the test set.

b. Excessively common words, or regular words such as conjunctions, prepositions and articles such as 'a' and 'of', may also be eliminated because of their high frequency of occurrence so as to not discriminate any specific class. Common words can be recognized either by a threshold on the number of documents the word occurs in, e.g. if it occurs in over half of all documents, or by supplying a *stop word* list. Stop word are language-specific and often domain-specific. Depending on the classification task, they may run the risk of removing words that are essential predictors, e.g. the word 'can' is discriminating between 'aluminum' and 'glass' recycling.

ii. Second in the process it is stated that the commonprocess of *stemming* or *lemmatizing—merging* various word forms such as plurals and verb conjugations into one distinct term—also reduces the number of features to be considered and which however is a feature engineering option. Suffix stripping Suffix stripping is used to stem words having a common stem and similar meanings can be merged into one term. Example, "invent," "invented," "inventing,""inventive," "invention," and "inventions" can be combined into the same term "invent" by removing the suffixes.

iii. Attribute selection. Other than the simple steps of stop-word removal and suffix stripping, attribute selection is a important step that can usually lessen significantly the attributes count. The attributes are typically term weights (determined by an indexing method) and the attribute space results in high computational overhead and increased training times making its removal necessary. It is depicted as a vector of features in a vector space model [5] or "bag-of-words" in a probabilistic model; features are the components in a vector or "words".

The filtering process is usually chosen because it is easily understood and has independent classifiers. In the filter approach, the attributes are evaluated based on some relevance measure, independent of any learning algorithm. In this paper, we use the term "relevance" informally to refer to the degree to which an attribute is relevant to the prediction of the class. For a proper definition of "relevance," please refer to Avrim and Pat (1997) [14]. The relevance measure is designed to measure the dependency between the class and an attribute and the attributes most applicable for predicting the class are selected. Since the attributes required to be evaluated only one time, the filter method is computationally efficient.

However, the attributes selected are not particularly trained on the learning algorithm used as it is actually not used in building the classifier. Also as the attributes are mostly individually assessed, the selected attributes, when considered as a set, may not be the most excellent possible subset.

In automatic classification, feature size reduction using simple filtering methods like stop words deletion or words stemming gives inadequate results. So a feature selection technique or algorithms ought to be used to optimize the performance of classification systems for visual detection.

2. The wrapper - The wrapper approach fundamentally depends on the learning algorithm. There are two major components in the wrapper approach: the (i) Performance Evaluation Method and (ii) Search Method.

   Cross-validation has been shown to be an effective performance evaluation. Cross- validation used to select feature generator is also used to tune the parameter automatically on the training data for selecting parameters for the induction algorithm, like the popular complexity constant C used in the SVM model. Optimizing each parameter in its own nested loop is the easiest to program, however, with each successive nesting a lesser fraction of the training data is provided to the induction algorithm. For instance if nested 5-fold cross-validation is used to decide on the feature generator, the number of features and also the complexity constant then the inner-most loop trains with only half of the training set: $\frac{4}{5}\times\frac{4}{5}\times\frac{4}{5}=51\%$. However the small training set fails in comparison to the full size training set for determining the optimal parameter values. So a 10-fold cross-validation, in spite of the high computing cost, is typically preferred to 2-fold cross-validation. As an alternative, a single loop of cross-validation must be combined with a multi parameter search strategy. The easiest way of programming is done by measuring the cross validation accuracy (or F-measure) at each point on a simple grid, and then deciding on the top parameters. There has been a huge research done on multi-parameter optimization and the methods though more complex to program are much more efficient. In the wrapper approach, the subset of features is chosen based on the Accuracy of classifiers. Exhaustively trying all
3. the subsets is not computationally feasible [15]. Technically, the wrapper is relatively difficult to implement, especially with a large amount of data.

   An optional feature selection process is the depiction of feature value. Usually for most of the cases it is adequate if a Boolean indicator for the word occurrence in the document. Additional options are; the number of times in the document the word occurs, the frequency of its incidence normalized by the length of the document, the count normalized by the inverse document frequency of the word. In cases where there are wide variations of document length, it can be essential to normalize the counts. In our study the datasets of most documents considered are short, that does not require any normalization. Also the words in short documents most probably do not repeat, resulting in Boolean word indicators to be as informative as counts. The result is of vast savings in training resources and in the search space of the induction algorithm. If not it may attempt to discrete each feature optimally, searching over the number of bins and each bin's threshold. In our study, we had chosen Boolean indicators for each feature that enlarges the selection of FS metrics that would be considered, e.g. Odds Ratio deals with Boolean features, and was reported by Mladenic and Grobelnik (1999) to perform well [16].
4. A final alternative in the FS strategy is if we can remove out all negatively correlated features. Some think that classifiers built from positive features alone will be efficient in the particular cases wherever the background class may shift and retraining is not required, which however has to be proved. Further, certain classifiers work basically with positive features, e.g. the Multinomial Naïve Bayes model and results prove it to be superior compared to previous Naïve Bayes model (McCallum & Nigam, 1998) [11], though significantly mediocre to some induction methods for text classification (e.g., Yang & Liu, 1999; Dumais et al., 1998) [17, 35]. Negative features are abundant due to the large class skew however rather important in practically: For example, while scanning a catalog of Web search results intended for the author's home page if many of results on George Foreman the boxer are shown they could be removed from the search with the terms 'boxer' and 'champion,' which is no concern to the author.

d) Feature Selection

i. *Prologue*

Feature selection refers to the selection of those features that are more important for relevant and informative data useful for text categorization. It enhances the scalability, efficiency as well as accuracy of a text classifier and plays a very important role in later steps influencing overall system performance. As many systems are large scale in various areas of data collection, feature selection is an important and widely grown. Some of basic applications of feature selection are Image Recognition, Clustering, Text Categorization, System monitoring, Rule Induction and Bioinformatics. (Jensen 2005) [48].

Feature selection consists after preprocessing and feature filtering selects the best features depending on the highest scores.

In document categorization or text classification, various methods of feature selection are used and they are;

Filter methods evaluate each feature independently and determine a ranking of all the features, from which the top ranked features are selected [4]. They can also be used as a pre-processing step to reduce the feature dimensionality to enable other, less scalable methods.

Wrapper methods search for the 'best' subset of features, repeatedly evaluating different feature subsets via cross validation with a particular induction algorithm. Wrapper methods have traditionally sought specific combinations of individual features from the power set of features, but this approach scales poorly for the large number of features inherent with classifying text. The wrapper methods have higher time complexity and accuracy compared to filter methods.

Embedded methods build a usually linear prediction model that tries to maximize the goodness-of-fit of the model and at the same time minimizes the number of input features [18].

Cross-validation method is used to select the best among feature generators and optimize other parameters, is somewhat like a wrapper method, but one that involves far fewer runs of the induction algorithm than typical wrapper feature selection.

Some variants build a classifier on the complete dataset where the classifier deletes the features it finds no application iteratively [19] as they are minimally scalable. However in case of large feature spaces, the memory may be insufficient for representing all the potential features and vectors. In this study such methods are not considered.

Text Classifiers Evaluation: Performance evaluation of the classifiers is the last stage of text classification. It is an experimental evaluation and not an analytical one. It is based on the capability and effectiveness of a classifier in taking the right categorization decisions rather than the Efficiency issues. The performance is measured with the help of many techniques like precision, recall [4], fallout, error, accuracy etc.;

(i) Precision w.r.t. ci (Pri) is defined as the as the probability that if a random document dx is classified under ci, this decision is correct.

(ii) Recall w.r.t. ci (Rei) is defined as the conditional that, if a random document dx ought to be classified under ci, this decision is taken, where TPi–The number of document correctly assigned to this category.

(iii) FN - The number of document incorrectly assigned to this category. FPi - The number of document incorrectly rejected assigned to this category. TNi -The number of document correctly rejected assigned to this category. Fallout = FNi / FNi + TNi

(iv) Error = FNi +FPi / TPi + FNi +FPi +TNi

(v) Accuracy = TPi + TNi

For obtaining estimates of precision and recall relative to the whole category set, methods such as i) Micro-averaging, ii) Macro-averaging are mostly used. Other measures like iii) Break–even point, iv) F-measure and v) Interpolation [7] are also used.

ii. *Feature Selection Methods and metrics*

The central design of Feature Selection (FS) is the selection of a subset of features from the original documents. In data mining or machine learning the methods that are regularly used for feature selection are: *Filter methods and Wrapper methods [12].*

Filtering methods: Filter methods use statistical techniques and are independent of the learning algorithm for FS. The filter method is usually chosen because it is easily understood and has independent classifiers. The various filtering metrics researched are; Document Frequency (DF), Term Frequency (TF-IDF), Chi Squared CHI2, Information Gain (IG), Accuracy (Acc2), Mutual Information (MI) [2], Association Word Mining [21], Expected Cross Entropy, Odds Ratio, Sampling Method, Gini Index etc. [22]. *The filter method is appropriate to treat very large feature space, is also the most scalable and is the focus of study in this paper;*

All features are evaluated independently with respect to the class labels in the training set to establish a ranking and the top ranking or scoring features are chosen [18] for classification. *A few filtering techniques metrics or scoring schemes are studied in this paper as they can be applied for most of the texts classification problems.* These filter metrics use a term goodness criterion threshold to attain a preferred degree of term purging in the entire terminology of a document. They are; (i) DF, (ii) TF-IDF (ii) Chi-square (iv) IG and (v) Acc2.

(i) DF: Document frequency is an easiest way of assessing feature significance; it simply determines in how many documents a word occurs where choosing regular words will advance the probability of the features presence in the next test cases. The DF of a specific term simply corresponds to the number of documents in a class containing that term [2, 4, 23]. It is computed independent of class labels and the total test set can also be included in the computation.

(ii) TF-IDF: Term Frequency method associates maximum scores to the terms that appear in some documents with a max frequency. That is a term occurring more number of times in a document means it is more discriminative whereas if it occurs in the majority of the documents, then it is less discriminative for the content.

(iii) Chi2Max: Chi-Squared is a statistical test that is widely used. It calculates the independence of 2 events between feature occurrence and class value [24] and the deviation from the expected distribution based on the assumption of actual independence.

(iv) IG: Information Gain measures in how much data the occurrence or nonexistence of a term or it

measures the decrease in entropy when the feature is given vs. absent that helps in deciding the correct classification choice for any class [2, 25]. IG reaches its highest value if a term is a perfect sign for class association, that is, if the term is occurs in a document and if and only if the document belongs to the respective class.

(v) Acc2: Accuracy considers only the number of documents in which the term occurs, without taking into account the number of actual documents.

Wrapper methods: Wrapper methods employ learning algorithm as the appraisal function. Classic AI search methods-such as simulated-annealing-to or greedy hill-climbing [19] explore for the 'best' subset of features and repetitively appraise different feature subsets with cross validation using a specific induction algorithm (Nejad et al., 2013) [20].

(i) Sequential Forward Selection (SFS),
(ii) Sequential Backward Selection (SBS),
(iii) Neural Networks (Dave, 2011, Eyheramendy and Madigan, 2005) [26],
(iv) Genetic Algorithm (GA) based selection.

The fourth method or Genetic Selection (GS) is a new FS employs the genetic algorithm (GA) optimization especially for issues of high dimensionality [3]. GA based selection has demonstrated to be reasonably capable and quick among many suboptimal search algorithms like sequential forward and backward selections [27]. GA theory is based on the survival of the fittest solutions from the entire potential solutions for a given issue [28]. Accordingly the latest generations formed from the surviving solutions are estimated to offer better accurateness to the best possible solution. The solutions match to chromosomes that are programmed with a proper alphabet. The fitness value of each chromosome is defined by a fitness function. New generations are generated by means of genetic operators i.e. crossover and mutation, with definite probabilities on the fittest members of the entire set. The primary set can be defined arbitrarily or manually. Population size, number of generations, probability of crossover and mutation are defined empirically. GS technique is simple and helpful and the chromosome length is equivalent to the dimension of a full feature set. The chromosomes are encoded as {0, 1} binary alphabet. In a chromosome, the indices denoted as "1" specify the chosen features, whereas "0" refers to the features not selected ones. For example, a chromosome defined as;

{ 1 0 1 0 1 1 0 0 0 1} implies that the 1st, 3rd, 5th, 6th, and 10th features are chosen and the remaining are eliminated. The fitness value related to a chromosome is defined by a specific success factor that is generated with the chosen features. A few instances of genetic feature selection research are presented in papers [27, 29, 30].

III. Contemporary Affirmation of the Literature About text Classification, Feature Selection and Filtering Strategies

a) *The Feature Selection Techniques or Methods*

This paper concentrates on filter methods because; i) they are comparatively more scalable to huge collections and ii) their objectives considerations are diverse from those of classifiers. There are four filter methods core, variant, combined and redundancy reducing methods;

i. *Core Methods*

We incorporated a few feature selection methods; (DF) document frequency (just count the number of documents including the feature), (IG) information gain (number of bits of information collected for category prediction for a particular feature) and (CHI) (measuring the absence of independence between a term and the category) [2]. Also the binary version of information gain (IG2) was included because it is widely used. Mutual information due to its poor performance was excluded.

ii. *Variant methods*

- Term frequency is used as a substitute for a binary value for every document counted in the scores (such variants would be recognized by TF in the results; because not one of these methods were amid the top three performers, they are not shown on the graphs.)
- The methods having one value per type (IG2, CHI, IG), we used average and also the maximum value as the score. (Identified by AVG, MAX)
- The methods, IG and CHI, were also tested with their generalized versions (cumulating evidence from all classes) are recognized by GEN.
- Also the rare words are eliminated (DF $\leq$ 5) (identified by "cut")

iii. *Combined methods*

We analyzed the correlation among some of the best performing methods and observed that a few (like the multiclass version of IG and CHI MAX) have minimum negative correlation, indicating a promising performance gain in combination. The two methods were combined by first normalizing the scores for all word and next selecting the higher of the two scores (thus giving an OR with equal weights to the two methods to be combined).

iv. *Redundancy Reducing Methods*

We executed a variant of the μ co-occurrence method as in [31] that utilizes the other filter feature

selection methods as a starting point. With the use of a tunable, arbitrary constant-size pool, the complexity analysis in [31] is seen to improve and we believe that using a percentage of the vocabulary is more suitable as the size of the vocabulary may vary broadly with collections. We executed a variant of this method, with a percentage-based initial pool (1% instead of 5 terms), smooth weighting in place of collection-dependent thresholding on the coocurrence and the multi-class version in place of 2-class.

As discussed above, the total number of features, variants as well as combinations is well over 100 where each of the core methods has an average of approximately 3 variants, and the cca. 15 resulting methods were combined in pairs.

b) Feature Selection Metrics - Evaluation and Exploration

In this study, we discuss criteria defining the feature selection metrics that have demonstrated excellent performance in text categorization. The five widely used feature selection metrics are: Document Frequency thresholding (DF), Term frequency-Inverse document frequency (TF-IDF), Chi-Square statistics (CHI), Information Gain (IG), Accuracy2 (Acc2);

i. *Document Frequency Threshold (DF)*

Document frequency is a metric used for remove the rare terms that are non-informative and confusing for classification. It is a very basic and accepted method that determines the number of documents in which the term appears without class labels [4, 32, 33]. It is based on the theory that a term belonging to a less number of documents is not an excellent feature for the classification task [34]. So, only the words that are present in a number of documents more than a defined threshold are chosen. This threshold can be calculated using a training set. Given a term t_k , this condition can be computed globally on the collection $(DFG(t_k))$ or on each category c_i $(DFL(t_k, c_i))$

$$DFL(t_k, c_i) = P(t \mid c) \approx \frac{A}{A+C}$$

$$DFL(t_k, c_i) = P(t \mid c) \approx \frac{A}{A+C}$$

One common technique to use this method is removing all the words which are present in less than x documents, x varying between 1 and 3 [34, 35, 36]. Commonly, this procedure is used with another feature selection method. The terms with low or high document frequency are frequently referred to as rare or common terms, in that order. The FS method discussed here is based on the first basic measurement that the terms with higher document frequency are more helpful for classification. However this supposition fails in giving any information sometimes. For instance the stop words (e.g., the, a, an) have very high DF scores, but hardly ever add to classification. More specifically, this uncomplicated method shoes good performance in few topic-based classification tasks (Yang and Pedersen, 1997).

ii. *Term Frequency-inverse Document Frequency (TF-IDF)*

The $tf - idf$ feature selection method is based on selecting the words with the highest $tf - idf$ scores. This method gives the highest scores to the words that are present in some documents with a high frequency meaning that it is more discriminative and if it is present in max number of the documents and then it is less discriminative for the content.

In $tf - idf$ [25], tf represents the term frequency of a term in a document. idf is defined as the inverse document frequency, i.e., the ratio of the total number of documents present in a dataset to the number of documents a given term appears in. A higher idf of a term implies that the term appears in relatively few documents and may be more significant at some stage in the process of text classification. *tfidf* is mostly used for term weighing in the field of information retrieval and is also used in text classification. The *tfidf* of a term t_k in document d_i is defined using;

$$tfidf(t_k, d_t) = tf(t_{,k}\, d_t) \log \frac{|D|}{df(t_k)}$$

Where $|D|$ refers to the total number of documents in a dataset; $tf(t_k, d_t)$ is the term frequency of a term t_k in document d_i; and $df(t_k)$ refers to the number of documents in which term t_k appears

iii. *Chi-square Statistics (CHI)*

Chi-square (2) statistics is a method commonly used in text categorization [4, 8, 32, 33], is a relevant measure, effective in text classification applications (Sebastiani 2002) [5] to measure the independence of two random variables (Liu and Setiono 1995) [37]. In text categorization, the two random variables are occurrence of term t_k and occurrence of class c_t and chi-square statistics measures the independence between t_k and c_t . The formula for chi-square score is:

$$CHI(t_k, c_t) = N \times \frac{[P(t_k, c_t)P(\overline{t_k}, \overline{c_t}) - P(\overline{t_k}, c_t)P(t_k, \overline{c_t})]^2}{P(t_k)P(\overline{t_k})P(c_t)P(\overline{c_t})}$$

where $P(t_k)$ is the percentage of documents in which term t_k occurs, $P(\overline{t_k})$ is the percentage of documents in which term t_k does not occur, $P(c_t)$ is the percentage of documents belonging to class c_t, $P(\overline{c_t})$ is the percentage of documents not belonging to class c_t, $P(t_k, c_t)$, is the percentage of documents belonging to class c_t in which term t_k occurs, $P(\overline{t_k}, \overline{c_i})$, is the percentage of documents not belonging to class c_t in which term t_k does not occur, $P(\overline{t_k}, c_i)$ is the percentage of documents belonging to class c_t in which term t_k does not occur and $P(t_k, \overline{c_t})$ is the percentage of documents not belonging to class c_t in which term t_k occurs. If chi-square score of a term t_k is of low value, this means t_k is independent from the class c_t and if chi-square score of a term t_k is of high value, this means t_k is dependent of the class c_t. Thus the chi-square feature selection method selects the terms with the highest chi-square score which are more informative for classification.

Due to the presence of words that rarely occurs and also due to limited number of positive training instances irregular behavior for very small expected counts, common in text classification, is observed.

iv. *Information Gain (IG)*

An accepted feature selection method in text categorization, information gain (IG) [4, 33, 38, 39] measures how much information the occurrence or nonexistence of a term helps to decide the correct classification criteria for any class [2,4, 40]. The terms with scores of highest information gain has maximum information about the classes. Here class membership and the presence/absence of a specific term in a certain category are seen as random variables; one computes how much information about the class membership is gained by knowing the presence/absence statistics. If the class membership is defined as a random variable c with two values, positive (c) and negative $(\bar{c})$, and a term is likewise seen as a random variable t with two values, present (t) and absent $(\bar{t})$, then information gain is calculated as;

$$IG(t_k, c_t) = \sum_{c \in [c_t, \overline{c_t}]} \sum_{t \in [t_k, \overline{t_k}]} P(t/c) \log \frac{P(t/c)}{P(t)P(c)}$$

5. Accuracy2 (Acc2)

Accuracy2 has showed better efficiency in comparison to other feature selection metrics in the earlier studies [4, 32]. In this metric, only the number of documents in which the term occurs is considered and not the number of actual documents. It measures the difference between the documents belonging to a class with a distributed term in the documents not belonging to t_k that class. Thus, the term that never occurs in a class c_t can be selected as a feature for c_t. Below is the formula for calculation of accuracy2 score:

$$Acc2(t_k, c_t) = \left| P(t_k, c_t) - P(t_k, \overline{c_t}) \right|$$

c) Feature Classification Strategies

The classification approaches for categorizing the selected features is performed by using 3 methods;

(i) Binary Classification,
(ii) Multi-Class Classification, and
(iii) Hierarchical Classification.

i. *Binary Classification*

Binary or binomial classification is categorizing the components of a given set into two sets based on specified classification rule. Binary domain tasks are regularly used and also as a subroutine to address maximum types of multi-class tasks.

Some usual binary classification tasks are (i) the effectiveness to the user in distinguishing spam email from good email. (ii) a "pass or fail" test method or quality control in factories; i.e. deciding if a specification has or has not been met: a Go/no go classification; (iii) an item may have a Qualitative property; it does or does not have a specified characteristic information retrieval, namely deciding whether a page or an article should be in the result set of a search or not – the classification property is the relevance of the article. (iv) to decide in medical testing if a patient has a specific disease or not – the classification property is the presence of the disease.

The two groups are not symmetric and this is observed in many practical binary classification instances. The focus is on the relative proportion of varied types of errors rather than on the overall accuracy. For example, in medical testing, a false positive (detecting a disease when it is not present) is considered differently from a false negative (not detecting a disease when it is present).

ii. *Multi-Class Classification*

There are two main types of multi-class classification:

(i) Single-label $(1-of-n)$ classification, where every individual case belongs to exactly one of the n classes. In the single-label case, many induction algorithms function by decomposing the problem into n binary tasks and then arriving at a final

decision by some sort of voting. Here also, feature selection can be optimized separately for each binary subtask. However, some $1-of-n$ induction algorithms do not execute binary decompositions, and require multi-class feature selection to choose a single set of features that perform well for the many classes.

(ii) Multi-label $(m-of-n)$ classification, where every individual case may belong to several, none, or even all classes. In the multi-label case, the difficulty is logically decomposed into n binary classification tasks: $class_i.vs.notclass_i$. These Individual binary tasks are solved independently where each may comprise its own feature selection to enhance its precision. And also a few $(m-of-n)$ applications programmable de novo require multi-class feature selection for performance and scalability reasons.

Other $(1-of-n)$ induction algorithms carry out a good deal of binary decomposition, e.g. algorithms finding optimal splitting hierarchies, or error-correcting code classifiers based on $o(n^2)$ dichotomies. In case of problems of this type, possibly we may carry out one multi-class feature selection rather than a separate binary feature selection for each dichotomy.

Theoretically all multi-class tasks could be performed with binary decompositions eliminating the requirement for multi-class feature selection. However, in reality lots of excellent software products APIs and libraries suppose the conversion of text into numerical feature vectors to be executed as a pre-processing step, and devoid of any capability for injecting feature selection into the inner loops, where the decompositions occur.

For instance, a centralized server task to classify millions of items on the network into multiple, orthogonal taxonomies, may be performed with more efficiency to establish a single, plausible sized feature vector to send through the network rather than sending individually to all the large documents.

For an application [41], of huge database of unstructured, multi-field (technical support) cases has memory by a cached, limited size feature vector representation for quick interactive examination, classification and labeling into multiple $(1-of-n)$ and $(m-of-n)$ taxonomies, where the classifiers are from time to time retrained in real time. It would be unfeasible to re-extract features for every binary decomposition or union of all the features into a exceptionally long feature vector that would be requested by all the binary feature selection sub tasks.

From various schemes of multi-class feature selection a few methods such as Chi-squared logically scale to multiple classes. However they face an underlying problem that is; an instance of a multi-class topic recognition case, with one of the classes holding all German texts. Now the German class will create many exceptionally predictive words. Almost all feature selection methods favor the stronger features and limit other classes for features. Similarly, if one class is mainly complicated, multi-class feature selectors will be inclined to disregard it, in view of the fact that it presents no strong features. These difficult classes require more features rather than fewer features.

A way out to this dilemma is to execute feature selection separately for each class through binary decompositions, and then to decide the final ranking of features using a round-robin algorithm where each class gets to vote its most preferred features in turn [41]. Since few classes are simpler to recognize than others this enhances performance even for well-balanced research benchmarks, however the difference results in most feature selection methods to be ignored, the very features that require most help. The aim of this scheme is to advance strength in atypical situations that arise only sporadically in practice that affects the average performance.

iii. *Hierarchical Classifications*

Hierarchy is one of the most predominant strategies for organizing abstractions. Hierarchical classification involves multiple tasks with the aim to classify items into a set of classes for organizing into a tree or directed acyclic graph, such as the Yahoo web directory. Here for some settings, the task is a single label problem to select $(1-of-n)$ nodes—or even limited to the leaf classes in the case of a 'virtual hierarchy.' For some other settings, the problem is of a multi-label task to select multiple interior nodes, optionally including all super-classes along the paths to the root.

Regardless of the given hierarchy of the classes, such issues are occasionally considered simply as flat multi-class tasks, either accommodating training examples up the tree structure for each class or a top-down hierarchy of classifiers may be generated to match the class hierarchy. The training set for each step down the tree is composed of all the training instances under each child subtree, optionally including a set of items positioned at the interior node itself, which terminates the recursion. Although this decomposition of classes is different from a flat treatment of the problem, in either decomposition, the same single-label or multi-label feature selection methods apply to the many sub-problems. It has been proposed that each internal hierarchical classifier may be quicker because dependency of each can be only for a few features (selected by feature selection), and can be further accurate because it only takes into account cases within a limited framework.

For instance an interior node concerning recycling with subtopics for glass recycling and can recycling there might be separate classifier intended for cases involving recycling. In this approach the training sets of every interior classifier are more balanced compared to a flat treatment of the problem.

iv. *Benchmarking Feature Selection Strategies*

The table 2.4 outlines latest research evaluating attributes selection techniques. The main research findings were,

(i) the filter method when implemented, information gain and chi-square have shown reasonably excellent performance (Yang and Pedersen (1997) [2] and the wrapper technique may have given even better performance if time had permitted in comparison to other similar measures.

(ii) Debole and Sebastiani (2003) [8] stated that gain ratio and chi-square outperformed information gain, iii) Forman (2003) [4] reported that information gain performed better than 10 other attribute selection methods in most experiments,

(iii) Halland Holmes 2003 [42] stated that wrapper method is very expensive for large datasets consisting of huge number of attributes.

Table 2.4 : Prior Studies on Attribute Selection Methods

References	attribute selection method	Outcome
Debole and Sebastiani 2003	χ^2, information gain, gr	GR and χ^2 outperformed Information Gain
Forman (2003)	accuracy, accuracy balanced, χ^2, document frequency, F1 measure, information gain, odds ratio numerator, odds ratio, pow, pr, rand	Information Gain outperformed other methods in most situations
Hall and Holmes (2003)	Correlation-based Feature Selection, information gain, wrapper, relief, consistency based, principle component	Wrapper was not applicable on the dataset with 1557 attributes due to time limitation. Wrapper and Correlation-Based Feature Selection outperformed other methods on the dataset with 293 attributes by NB.
Lewis and Ringuette	information gain	Both propBayes and DT-MIN 10 provided reasonable performance
Liu(2004)	information gain, mutual information, χ^2, odds ratio, simplified-chi-square	Information Gain an d χ^2 were most effective for NB. No benefit for SVM was found.
McCallum and Nigam (1998)	information gain	MNB outperformed NB in large attributes set
Madenic (1994)	information gain, odds ratio, word frequency, rand	Odds Ratio outperformed other methods
Ribone (2002)	information gain, word frequency, document frequency	Information Gain>Word Frequency>Document Frequency
Rogati and Yang (2002)	Document frequency, information gain χ^2	χ^2 outperformed other methods
Joachims (1996)	information gain	SVM outperformed other classifiers
Liu(2002)	χ^2, Correlation-based Feature Selection, mit correlation, entropy,	Entropy was the best, followed by χ^2, on the datasets. Correlation-based feature selection outperformed others on the ovarian cancer dataset.
Sebastiani (2002)		Summary of previous studies as {Odds Ratio, NGL coefficient, SS}>{ χ^2, Information Gain}>Mutual Information
Yang and Pedersen (1997)	Document frequency, information gain, mutual information, χ^2, term strength	Information Gain and χ^2 were most effective. Performance improved after attribute selection

Note 1: Abbreviations of attribute selection methods: ACC—Accuracy; ACC2—Accuracy balanced; BNS—Bi-Normal Separation; CFS—Correlation-based Feature Selection; χ2—chi-square; CNS Consistency-based; DF—document frequency; F1—F1 Measure; GSS—GSS coefficient (simplified chi-square); IG—information gain; MI—mutual information; NGL—NGL coefficient; ODDN—odds ratio numerator; OR—odds ratios; PC—Principal Components; POW—Power; PR—Probability Ratio; RAND—Random; RLF—Relief; TS—term strength; WF—word frequency; WRP—Wrapper.

Note 2: Abbreviations of classification methods: C4.5—decision tree; DT-min10—decision tree; kNN—k-Nearest Neighbors; LLSF—Linear Least Squares Fit; LR—Logistic Regression; NN—Neural Network; NB—Naïve Bayes; MNB—Multinomial Naïve Bayes; PCL—Prediction by Collective Likelihood; Prop Bayes—Bayesian classifier; SVM—Support Vector Machine.
Note 3: "$>$" means "performed better than".

d) Combination of Feature Selection Methods

Several researches have been done to enhance the efficiency of feature selection strategies on text categorization; however they generally are about improving the performance of the individual feature selection methods. The success of a feature selection method is determined by various variables and it is very difficult to understand which method is better performer than others though several selection methods are prevalent. So the combination of distinct feature selection methodologies gives more efficiency in text classification. The output of classifiers combination is a potential strategy and it is being studied extensively in the area of information retrieval. This section covers strategies for combining the outputs of the individual feature selections metrics.

i. *The Common Combining Strategies*

There are several strategies of combining the outputs of the different FS methods. Some combination strategies or most popular approaches to combine various outputs are;

1. Linear Combining Methods such as
   (i) Averaging and
   (ii) Weighted Averaging
2. Non-Linear Combining Methods such as
   (i) Ranking and
   (ii) Voting

Averaging (Tumer and Ghosh, 1999) [43], a Linear Combining method is the most frequently used combining strategy. Fox and Shaw, 1994, significantly state that most excellent combining strategy depends on adding the outputs of the algorithms that is similar to averaging on comparing six combining strategies [44]. Also Hull et al., 1996, [45] show that the accuracy of the simple averaging strategy is far better than the complex combinations of the classifiers.

With respect to the above studies and considering the results, we decide to apply the averaging strategy in two ways: *(1-a) Score Combination and (1-b) Rank Combination.*

In the Score and Rank Combination Strategy, the preceding research finds that the efficiency of the combination and the number of feature selection methods considered in combination are inversely related and vice versa. Also it is reported that maximum efficiency is achieved essentially by the combination of two feature selection methods [33, 46]. So we only choose combining two distinct feature selection methods. In the paper, we assess the performance of all potential binary-combinations (2-combinations) of five different feature selection methods: DF, TF-IDF, CHI, IG, Acc2 and metrics which are TF-IDF & CHI, TF-IDF & IG, TF-IDF & Acc2, TF-IDF & DF, CHI & IG, CHI & Acc2, CHI & DF, IG & Acc2, IG & DF and Acc2& DF.

The strategy of the feature selection methods is of two steps. In the first step score is given to the terms that are more relevant for classification and in the second step is of selecting the terms with highest score. In scoring stage, the different feature selection methods and their scores of each term are normalized using the maximum and minimum scores according to the below formula:

Score $= (s1, s2, \ldots, sn)$ where si score of the i^{th} term, n is the total number of term.

$$score^{1} = \frac{Score - \min(Score)}{\max(Score) - \min(Score)}$$

By normalization, the scores fall in the same range [0, 1] and scores of the terms from the different feature selection methods are represented equally which facilitates efficient comparisons between the methods.

(1-a) Score Combination

Score combination is averaging the normalized term scores of the different feature selection methods.

$$c_{score} = \sum_{i-1}^{M} \frac{Score_i^{'}}{M}$$

where M is the number of feature selection methods which is 2 for this study.

(1-b) Rank Combination

Rank Combination is averaging the term ranks collected from the term scores of the different feature selection methods.

Rank $= (r1, r2, \ldots, rn)$ where ri rank of the i^{th} term, n is the total number of term.

$$c_{rank} = \sum_{i=1}^{M} \frac{Rank_i}{M}$$

There are several modes for determining rankings like i) Standard Competition Ranking, ii) Modified Competition Ranking, iii) Dense Ranking, iv) Ordinal Ranking and v) Fractional Ranking. We rank the terms in our research as per the descending order of their scores and with standard competition ranking strategy. Here in competition ranking ("1, 2, 2, 4" ranking), terms with similar score are assigned the same ranking number and next a gap is given in the ranking numbers.

ii. *Proposed Combinations of Feature Selection Methods*

We present seven latest methods for the efficiency in combining the different metrics for feature selection and compare them as discussed below;

In the seven proposed combinations the first and second combinations are comparable to score combination and are a variation of the score combination. The remaining five combinations are a product combination of both score and rank value of the terms.

a. *C1 Combination -- Logarithmic Combination*

The first method, the logarithmic combination is based on the score combination that is simply averaging the term scores of the feature selection methods. The principle applied is the same where in the proposed combination the logarithmic scores of the terms are averaged in place of term scores. The principal involved is increasing the interval between the highest and the lowest scores, taking benefit of the logarithm. The interval between the scores increases as the scores decrease and it decreases as the scores increase. The natural logarithm is used as a logarithm function where if the value of the score is zero, then it will substitute the value with 0.00001 for computing the logarithm. The calculation of the C1 Combination is given by the following formula:

$$C_1 = \sum_{i=1}^{M} \frac{\ln(score_i^{'})}{M}$$

b. *C2 Combination – Square Combination*

In the second method, the Square Combination proposed is also based on the score combination, however the term scores are squared and averaged rather than only averaging the term scores. The interval between the scores exponentially increases rather than the interval between the highest scores and lowest scores that remains unchanged with the increases of scores. The difference between the scores increases as the scores increase and decreases as the scores decrease in contrast to preceding methods. Thus, if the term is considered essential for classification, this method elevates the importance of the term compared to the remaining terms prior to combination. The formula used is:

$$C_2 = \sum_{i=1}^{M} \frac{(Score_i^{'})^2}{M}$$

c. *C3 Combination – Product Combination with Fraction*

The third method, the product combination with fraction (C3) initially the rank of the term is multiplied with the scores of the terms. Next the outputs of the multiplication of the individual feature selection methods are added and divided with one for combining the outputs of the feature selection metrics. In case the value of score is equal to 1, then it will substitute the value with 0.99999 to prevent division by zero. The highest value terms are selected in feature selection step. The formula is:

$$C_3 = \frac{1}{\sum_{i=1}^{M} Rank_i^{x}(1 - Score_i^{'})}$$

When we assess the effectiveness of the combinations in classification we see that both score and rank combinations of the feature selection methods enhance the performance of the individual methods. This led us to consider using both rank and score values in the same function. The central idea of the "product combinations" in 1, 2, 3, 4, 5, 6 and 7 is to take benefit of the score and rank of the terms while identifying the most effective terms. Here both rank of the term and the score of the term have equal weight in the combination. Based on this principle, if the scores of terms of different feature selection methods are equal then greater importance is given to the term with highest rank. Different versions of the product combination are derived and are described below.

d. *C4 Combination – Product Combination with Fraction*

The fourth method, the product combination with fraction (C4) is different from the above proposed third method. In this method the logarithm of the rank is multiplied with the square of the score of the term. The formula of the C4 Combination is given as below:

$$C_4 = \frac{1}{\sum_{i=1}^{M} \ln(Rank_i) x (1 - Score_i^{'})^2}$$

e. *C5 Combination - Product Combination with Fraction*

The fifth method, the product combination with fraction, is different from the third method discussed above. In this method the square root of the rank is multiplied with the square of the score of the term. The formula is:

$$C_5 = \frac{1}{\sum_{i=1}^{M} (Rank_i)^{1/2} \times (1 - Score_i^{'})^2}$$

f. *C6 Combination - Logarithmic Product Combination*

The sixth method, the logarithmic product combination is sum of the products of each terms logarithm of the rank and the logarithm of the score of each feature selection method. The formula is:

$$C_6 = \sum_{i=1}^{M} \ln(Rank_i) x \ln(Score_i^{'})$$

g. *C7 Combination - Product Combination*

The seventh method, the product combination or C7 Combination is the last method that is the multiplication of the square root of the rank with the logarithm of the score of the term and addition of the outputs. The formula is:

$$C_7 = \sum_{i=1}^{M} (Rank_i)^{1/2} x \ln(Score_i')$$

IV. Conclusion

Feature Selection [47] continually handles huge and complex datasets and faces typical problems of feature space, which if very high, increases computational costs and also the training time.

Feature selection relevance has decreased with the constant developments in accuracy and scalability of core machine learning algorithms. An algorithm for instance, when considering more than 800,000 text cases was developed by Joachims that is an innovative linear SVM classifier. Using a 3.6GHz PC processor [18], it may be trained with approximately 50,000 word features in less than 3 minutes. As in some cases where feature selection does not improve the accuracy of the training sets, researchers interested in approaches other than feature selection can avoid the problems associated with feature selection and go for an input representation that is fixed and conveniently replicable.

However a case where a researcher of data mining is provided with a training set for generating an excellent possible classifier should definitely consider feature selection. The method can improve certainly Accuracy for certain datasets or at least give slight improvements on average. Thus, feature selection still has a role to play for those who seek to maximize Accuracy, e.g. industrial practitioners, application programmers and contestants in data-mining competitions.

References Références Referencias

1. F. Sebastiani, "Text categorization", Alessandro Zanasi (ed.) Text Mining and its Applications, WIT Press, Southampton, UK, pp. 109-129, 2005.
2. Wei Zhao "A New Feature Selection Algorithm in Text Categorization "International Symposium on Computer, Communication, Control and Automation 2010.
3. Y. Yang, J.O. Pedersen, "A comparative study on feature selection in text categorization", Proceedings of the 14th International Conference on Machine Learning, pp. 412–420, 1997.
4. G. Forman, "An extensive empirical study of feature selection metrics for text classification", Journal of Machine Learning Research, Vol. 3, pp. 1289–1305, 2003
5. Sebastiani F (2002). Machine Learning in Automated Text Categorization, ACM Computing Surveys, 34(1): 1-47.
6. Jensen R (2005). Combining rough and fuzzy sets for feature selection, PhD Thesis, University of Edinburgh, UK.
7. KjerstiAas and Line Eikvil "Text Categorization: A Survey" Report No. 941. ISBN 82-539-0425-8. , June, 1999.
8. Debole, F. And F. Sebastiani, "Supervised Term Weighting for Automated Text Categorization", Proceedings of SAC-03-18th ACM Symposium on Applied Computing, ACM Press, pp. 784–788, 2003.
9. Wanas, N., D. A. Said, N. H. Hegazy and N. M. Darwish, "A Study of Local and Global Thresholding Techniques in Text Categorization", Proc. Fifth Australasian Data Mining Conference, 2006.
10. D. Mladenic and M. Globelnik. Word sequences as features in text learning. In Proceedings of the 17th Electrotechnical and Computer Sci-ence Conference (ERK98), Ljubljana, Slovenia, pages 145–148, 1998.
11. A. McCallum and K. Nigam. A comparison of event models for naïve bayes text classification. In AAAI/ICML-98 Workshop on Learning for Text Categorization, TR WS-98-05, pages 41–48. AAAI Press, 1998.
12. S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In International Conference on Machine Learning, 2001.
13. George A. Miller and Edwin B. Newman. Tests of a statistical explanation of the rank-frequency relation for words in written English. American Journal of Psychology, 71:209-218, 1958.
14. Avrim, L.B., and Pat, L. "Selection of Relevant Features and Examples in Machine Learning.," Artificial Intelligence (97:1-2) 1997, pp 245-271.
15. D. Koller and M. Sahami. Toward optimal feature selection. In International Conference on Machine Learning, pages 284–292, 1996.
16. DunjaMladenic and Marko Grobelnik. Feature Selection for Unbalanced Class Distribution and Naïve Bayes. In Proceedings of the Sixteenth International Conference on Machine Learning (ICML), pages 258-267, 1999.
17. Yiming Yang and Xin Liu. A Re-examination of Text Categorization Methods. In Proceedings of the Twenty-Second International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 42-49, 1999.
18. T. Joachims. Training linear SVMs in linear time. In Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 217–226, 2006.
19. I. Guyon and E. Elisseef, A. Special issue on variable and feature selection. J. of Machine Learning Research, 3:1157–1461, 2003.
20. Nejad MB, Attarzadeh I, Hosseinzadeh M (2013). An Efficient Method for Automatic Text Categorization",

International Journal of Mechatronics, Electrical and Computer Technology, 3(9): 314-329.

21. Su-JeongKo and Jung-Hyun Lee "Feature Selection Using Association Word Mining for Classification "H.C. Mayr et al. (Eds.): DEXA 2001, LNCS 2113, pp. 211–220, 2001.
22. AnirbanDasgupta "Feature Selection Methods for Text Classification "KDD'07, August 12–15, 2007.
23. C.D. Manning, P. Raghavan, H. Sch¨utze, Introduction to Information Retrieval, Cambridge, Cambridge University Press, 2008.
24. Yang,Y. (1999).An evaluation of statistical approaches to text categorization. Information Retrieval, 1(1–2), 69–90.
25. Chih, H.,&Kulathuramaiyer,N. (2004). An empirical study of feature selection for text Categorization based on term weightage. In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (pp. 599–602).Washington, DC: IEEE.
26. Dave K (2011). Study of feature selection algorithms for text-categorization, University of Nevada, Las Vegas, UNLV Theses/Dissertations/Professional Papers/ Capstones, Paper 1380.
27. S. Gunal, O.N. Gerek, D.G. Ece, R. Edizkan, "The search for optimal feature set in power quality event classification", Expert Systems with Applications, Vol. 36, pp. 10266–10273, 2009.
28. D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Reading, Massachusetts, Addison-Wesley, 1989.
29. C.L. Huang, C.J. Wang, "A GA-based feature selection and parameters optimization for support vector machines", Expert Systems with Applications, Vol. 31, pp. 231–240, 2006.
30. J. Yang, V. Honavar, "Feature subset selection using a genetic algorithm", IEEE Intelligent Systems, Vol. 13, pp. 44–49, 1998
31. P. Soucy and P. Mineau. A simple feature selection method for text classification. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, pages 897–902, 2001.
32. Tasci, S., "An evaluation of existing and new feature selection metrics in text categorization", Computer Engineering, Bogaziçi University, 2006
33. Li, Y., D. F. Hsu and S. M. Chung "Combining Multiple Feature Selection Methods for Text Categorization by Using Rank-Score Characteristics", International Conference on Tools with Artificial Intelligence - ICTAI, pp. 508-517, 2009
34. E. Wiener, J. O. Pedersen, and A. S. Weigend. A neural network approach to topic spotting. In SDAIR'95: Proceedings of the 4th Symposium on Document Analysis and Information Retrieval, pages 317–332, 1995.
35. S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In CIKM'98: Proceedings of the 7th international conference on Information and knowledge management, pages 148–155, New York, NY, USA, 1998. ACM.
36. Y. H. Li and A. K. Jain. Classification of text documents. The Computer Journal, 41:537–546, 1998.
37. Liu, H., and Setiono, R. "Chi2: Feature Selection and Discretization of Numeric Attributes," proceedings of the Seventh International Conference on Tools with Artificial Intelligence, 1995.
38. M. F. Caropreso, S. Matwin, and F. Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In A. G. Chin, editor, Text Databases and Document Management: Theory and Practice, pages 78–102. Idea Group Publishing, Hershey, US, 2001.
39. C. Lee, G.G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization", Information Processing and Management, Vol. 42, pp. 155–165, 2006.
40. Quinlan, J. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
41. G. Forman. A pitfall and solution in multi-class feature selection for text classification. In ICML '04: Proc. of the 21st Int'l Conf. on Machine learning, pages 297–304. ACM Press, 2004.
42. Hall, M.A., and Holmes, G. "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining," IEEE Transactions on Knowledge and Data Engineering (15:6) 2003, pp 1437-1447.
43. Tumer, K. and J. Ghosh, "Linear and order statistics combiners for pattern classification", Combining Artificial Neural Networks, Springer Verlag, pp. 127–162, 1999.
44. Fox, E. and J. Shaw, "Combination of multiple searches", Proceedings of the 2nd Text Retrieval Conference (TREC-2), National Institute of Standards and Technology Special Publication 500-215, pp. 243–252, 1994
45. Hull, D., J.Pedersen and H. Schutze, "Method combination for document filtering", Proceedings of the 19th ACM SIGIR Conference on Researchand Development in Information Retrieval, pp.279–287, 1996.
46. Olsson, J. S. and D. W. Oard, "Combining Feature Selectors for Text Classification", CIKM'06, 2006.
47. A. Dasgupta, P. Drineas, B. Harb, "Feature Selection Methods for Text Classification", KDD'07, ACM, 2007.
48. Jensen R (2005) Combining rough and fuzzy sets for feature selection, PhD thesis for University of Edinburgh UK.

Global Journals Inc. (US) Guidelines Handbook 2014

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

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

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

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

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

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

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

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

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

It is mandatory to read all terms and conditions carefully.

AUXILIARY MEMBERSHIPS

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

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

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

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

The Institute will be entitled to following benefits:

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

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

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

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

The Board can also play vital role by exploring and giving valuable suggestions regarding the Standards of "Open Association of Research Society, U.S.A (OARS)" so that proper amendment can take place for the benefit of entire research community. We shall provide details of particular standard only on receipt of request from the Board.

The board members can also join us as Individual Fellow with 40% discount on total fees applicable to Individual Fellow. They will be entitled to avail all the benefits as declared. Please visit Individual Fellow-sub menu of GlobalJournals.org to have more relevant details.

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

After nomination of your institution as "Institutional Fellow" and constantly functioning successfully for one year, we can consider giving recognition to your institute to function as Regional/Zonal office on our behalf.
The board can also take up the additional allied activities for betterment after our consultation.

The following entitlements are applicable to individual Fellows:

Open Association of Research Society, U.S.A (OARS) By-laws states that an individual Fellow may use the designations as applicable, or the corresponding initials. The Credentials of individual Fellow and Associate designations signify that the individual has gained knowledge of the fundamental concepts. One is magnanimous and proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice.

Open Association of Research Society (US)/ Global Journals Incorporation (USA), as described in Corporate Statements, are educational, research publishing and professional membership organizations. Achieving our individual Fellow or Associate status is based mainly on meeting stated educational research requirements.

Disbursement of 40% Royalty earned through Global Journals : Researcher = 50%, Peer Reviewer = 37.50%, Institution = 12.50% E.g. Out of 40%, the 20% benefit should be passed on to researcher, 15 % benefit towards remuneration should be given to a reviewer and remaining 5% is to be retained by the institution.

We shall provide print version of 12 issues of any three journals [as per your requirement] out of our 38 journals worth $ 2376 USD.

Other:

The individual Fellow and Associate designations accredited by Open Association of Research Society (US) credentials signify guarantees following achievements:

- The professional accredited with Fellow honor, is entitled to various benefits viz. name, fame, honor, regular flow of income, secured bright future, social status etc.

- In addition to above, if one is single author, then entitled to 40% discount on publishing research paper and can get 10%discount if one is co-author or main author among group of authors.
- The Fellow can organize symposium/seminar/conference on behalf of Global Journals Incorporation (USA) and he/she can also attend the same organized by other institutes on behalf of Global Journals.
- The Fellow can become member of Editorial Board Member after completing 3yrs.
- The Fellow can earn 60% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.
- Fellow can also join as paid peer reviewer and earn 15% remuneration of author charges and can also get an opportunity to join as member of the Editorial Board of Global Journals Incorporation (USA)
- • This individual has learned the basic methods of applying those concepts and techniques to common challenging situations. This individual has further demonstrated an in–depth understanding of the application of suitable techniques to a particular area of research practice.

Note :

"

- In future, if the board feels the necessity to change any board member, the same can be done with the consent of the chairperson along with anyone board member without our approval.
- In case, the chairperson needs to be replaced then consent of 2/3rd board members are required and they are also required to jointly pass the resolution copy of which should be sent to us. In such case, it will be compulsory to obtain our approval before replacement.
- In case of "Difference of Opinion [if any]" among the Board members, our decision will be final and binding to everyone.

"

Process of submission of Research Paper

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

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

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

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Swis 721 Lt BT.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, "Abstract" word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be three lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications.

Research letters: The letters are small and concise comments on previously published matters.

5. STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

You must strictly follow above Author Guidelines before submitting your paper or else we will not at all be responsible for any corrections in future in any of the way.

Before start writing a good quality Computer Science Research Paper, let us first understand what is Computer Science Research Paper? So, Computer Science Research Paper is the paper which is written by professionals or scientists who are associated to Computer Science and Information Technology, or doing research study in these areas. If you are novel to this field then you can consult about this field from your supervisor or guide.

TECHNIQUES FOR WRITING A GOOD QUALITY RESEARCH PAPER:

1. Choosing the topic: In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. Pick a good study spot: To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. Think technically: Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page
- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.

- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.
- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

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

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