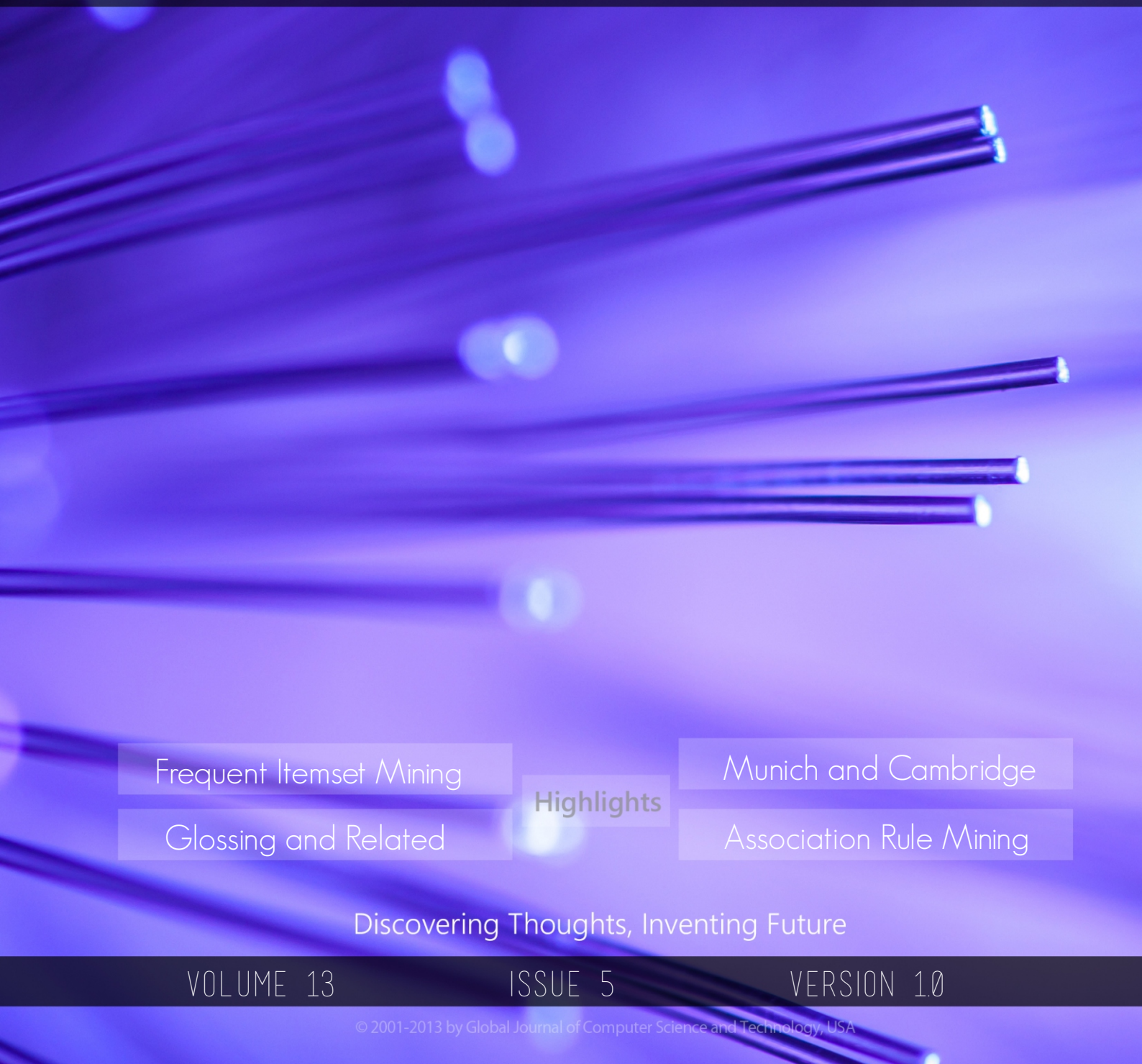
Online ISSN : 0975-4172
Print ISSN : 0975-4350

Global Journal

of Computer Science and Technology: C

Software & Data Engineering



Frequent Itemset Mining

Glossing and Related

Highlights

Munich and Cambridge

Association Rule Mining

Discovering Thoughts, Inventing Future

VOLUME 13 ISSUE 5 VERSION 1.0

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

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

Volume 13 Issue 5 (Ver. 1.0)

Open Association of Research Society

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

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

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

Publisher's Headquarters office

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

Offset Typesetting

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

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

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

eContacts

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

Pricing (Including by Air Parcel Charges):

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

Editorial Board Members (HON.)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Contents of the Volume

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 5 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Benchmark Algorithms and Models of Frequent Itemset Mining over Data Streams: Contemporary Affirmation of State of Art

By V. Sidda Reddy, Dr. T.V. Rao & Dr. A. Govardhan
K.L. University, India

Abstract - Data mining and knowledge discovery is an active research work and getting popular by the day because it can be applied in different type of data like web click streams, sensor networks, stock exchange data and time-series data and so on. Data streams are not devoid of research problems. This is attributed to non-stop data arrival in numerous, swift, varying with time, erratic and unrestricted data field. It is highly important to find the regular prototype in single pass data stream or minor number of passes when making use of limited space of memory. In this survey the review on the final progress in the study of regular model mining in data streams. Mining algorithms are talked about at length and further research directions have been suggested.

Keywords : *data stream mining; sliding window; training model; linear reparability; data mining, frequent pattern; combinatorial approximation; single-pass algorithms; bit-sequence representation.*

GJCST-C Classification : H.2.8

Strictly as per the compliance and regulations of:

Benchmark Algorithms and Models of Frequent Itemset Mining over Data Streams: Contemporary Affirmation of State of Art

V. Sidda Reddy [α], Dr. T.V. Rao [σ] & Dr. A. Govardhan [ρ]

Abstract - Data mining and knowledge discovery is an active research work and getting popular by the day because it can be applied in different type of data like web click streams, sensor networks, stock exchange data and time-series data and so on. Data streams are not devoid of research problems. This is attributed to non-stop data arrival in numerous, swift, varying with time, erratic and unrestricted data field. It is highly important to find the regular prototype in single pass data stream or minor number of passes when making use of limited space of memory. In this survey the review on the final progress in the study of regular model mining in data streams. Mining algorithms are talked about at length and further research directions have been suggested.

Keywords : data stream mining; sliding window; training model; linear reparability; data mining, frequent pattern; combinatorial approximation; single-pass algorithms; bit-sequence representation.

I. Introduction

Frequent item set mining [48] is popularly known to be very essential in several crucial data mining activities. These activities include clusters [20], classifiers [46], sequences [62], correlations [14] and associations [35]. Several studies that point to mining frequent item sets on statistic databases and several proficient algorithms are suggested.

In recent years, data streams become an active research work in computer applications such as database systems, distributed databases and data mining. Data streams also importance to study online mining of frequent item sets, which is much needed for

These applications include sensor networks, e-business and stock market analysis, trend analysis, fraud detection in telecommunications, network traffic analysis, and web log and click-stream mining. It is ever more tedious to perform data mining and advanced analysis on huge and rapidly arriving data streams to capture remarkable development, model and exceptions. This is recognized to the fast appearance of these new application domains.

Author α : Research scholar, JNTU, Hyderabad, AP, India.
E-mail : siddareddy.v@gmail.com
Author σ : Professor, School of Computing, K.L. University, AP, India.
E-mail : tv_venkat@yahoo.com
Author ρ : Professor and DE, JNTU, Hyderabad, AP, India.
E-mail : govardhan_cse@yahoo.co.in

Data streams are continuous and high speed flow of data items that come in an appropriate order. These are quite different when compared to the data in traditional static databases.

Apart from having data distributions that modify with time, the data streams are unbounded, usually come in high speed and are continuous [23].

Data stream is further divided into offline streams and online streams. Normal bulk arrivals are attributed to offline streams [13]. Making reports on web log streams is regarded as mining offline data streams this is due to the reports that are created on the log data for a specific period of time. Backup devices or queries on updates to warehouse form other examples for offline streams. Queries on these streams can be permitted to treated offline.

Online streams are distinguished by instantaneously restructured data that appear once at a time. Calculating the regularity estimation of the internet packet streams is an application of mining online data streams, since internet packet streams is an instant one at a time packet process. Sensor data, network measurements and stock tickers are the other online data streams. Apart from keeping pace with the high speed of online queries, these must also be developed online. Immediately on their arrival, they need to be processed. It is not possible to process bulk data for mining data streams comes with a new set of questions. Primarily, to keep the whole stream in the main memory or in a secondary storage area it will be impractical. This is because data is streamed non-stop and the quantity of data is limitless. Secondarily it is not feasible to use traditional mining techniques with multiple scans like stored datasets. In data streams, data will be streamed and flow with high speed only once. To keep pace with high data arrival rate the mining streams will need quick, real-time processing. The results will likely to be available in a short time span. Also when the item sets are combine it can intensify mining regular item sets over streams in regards to both processing frequency and memory consumption.

Due to these limitations, research work performed on approximating mining results, with some sensible assurance on the value of approximation.

a) Distinguish between data streams and static data

Static data contains persistent data relations, approach of querying is one-time, data set is passive and randomly accessed, no role of response time on result accuracy, update speed is minimal and responses in passive mode.

Data streams contains transient data relations, the querying process is continuous, data set with continuous updates and sequentially accessed, response time influences the performance and results accuracy, update speed is maximal since response type is active

II. Taxonomy of Issues in Mining Data Streams

Regular item set issue is a daunting problem. For instance, there can be a massive number of regular item sets because of combinational explosion. The test here lies in how effectively can we detail, find and stock up these regular item sets. Due to their exclusive features, these data streams have added many more new challenges in regular item set mining.

Lately, in a couple of data sources, the data generation has become rapid than before. This quick production of non-stop data streams has questioned storage, communication and computation capacity in the computing systems. It is quite demanding to store, mine and query these data sets. Pulling out knowledge structures present in models and patterns in the uncontrollable streams of information is what mining data streams are all about.

The demands and the research problems faced in the data stream mining is motivational. Several demands concerned with this arena are explained here.

a) Handling the continuous flow of data

This problem is concerning the data management. These high data rates cannot be handled by the traditional database management systems. As a result, it is unreasonable to keep all the information in relentless media. In addition, it is very high-priced to aimlessly check the data several times. The demanding task here is to access the data item once and get frequent item sets. To deal with these fluctuations it is important to see to new indexing, storage and querying techniques.

b) Data reduction and synopsis construction

To comply with the earlier mentioned problems the data stream analysis, classification, querying and clustering applications will need some type of specification techniques. These techniques will be used to give out fairly accurate answers from huge data sets generally by means of synopsis construction and data reduction. This can be done by choosing a subset of incoming data or by making use of sketching, aggregation techniques, load shedding.

c) Minimizing energy consumption

In the resource-constrained environment a massive quantity of data streams are produced. For example: Sensor network. Devices here have short battery life. Energy efficient designs are very important because sending every generated stream to a central site is energy inefficient in addition to its lack of scalability problem.

d) Unbounded memory requirements

Data streams have unbound data but the storage that can be used to find or preserve frequent item sets is inadequate. Machine learning techniques show the core basis of data mining algorithms. When the analysis algorithm is implemented, it is important that the machine learning methods have the information in the memory. Because of this large number of produced streams, it is extremely essential to design space effective techniques that can have one look or less over the incoming stream. The frequency of an item set depends on time; this is another result of unbounded data. It is very demanding to use inadequate storage and find dynamic frequent item sets from unbounded data.

e) Transferring data mining results

Representation of knowledge is a new important research essential. The structures must be transferred to the user after getting it from models and patterns from data stream generators. Kargupta et al [6] have dealt with this issue. He has used Fourier transformations to capably transfer the mining results in a limited bandwidth links.

f) Modeling changes of result over time

In a couple of cases, the user is disinterested in the mining data stream results. But how do these results change over time. For instance, clusters formed by changes in data stream, may symbolize the changes in the dynamics of the coming stream. This change would help many temporal-based analysis applications like disaster recovery, emergency and video-based surveillance.

g) Real-time response

There is a demand on response time as the data stream applications are very time specific. Algorithms that come slower than the data arriving rate in constrained situations are of no use.

h) Visualization of data mining results

Research is still on in revelation of traditional data mining results on a desktop. It is a real challenge to see visualization in the small screens of the PDA. If a businessman is seeing the results of data being streamed and analyzed on his PDA, the results should be so effective that it should facilitate him to take quick decisions.

III. Taxonomy of Data Preprocessing in Mining Data Streams

a) Sampling

Data size can be lessened by using the random sampling technique. This can be done while capturing its important attributes. By using this form of summarization in a data stream and other summation can be built by using this example itself [2]. The data size is mandatory to get unbiased sample of data. The sampling approach is the simplest approach that has periodic time intervals which gives nice way to dwindle the data stream. This approach can have huge information loss in the stream as the data rates change.

b) Sketching

Sketching is a process of constructing a statistical synopsis of a data stream that makes use a little amount of memory and has frequency moments.

c) Histograms

Histograms are synopsis structures that can total the distribution value of the dataset. Tasks such as data mining, approximate query answering and query size estimation uses histograms.

d) Wavelets

These are methods for estimating data with a known likelihood. Wavelets co-efficient are predictions of a known signal (set of data values) into an orthogonal set of source vectors.

e) Concept Drifts

This is a crucial field in data mining. We need to study the judgment of swift and precise concept drift, the successful use of concept drift acquisition, how to save and serious use of concept and inclination of concept drift.

f) Sliding Window Models

The recent most data streams are examined in the sliding window models. The recent data items and summarized versions are examined in detail. Many techniques have accepted by many techniques. A preset newly generated data items, intended for data mining are taken up and knowledge discovery is performed on this. To make this synopsis structure several other synopsis structures are conversed to the entire data stream. It is assumed that the sliding window model is inspired to use the most recent data in the data stream [9]. Hence only a set history of the data stream is taken into consideration for the analysis and processing.

g) Landmark-window models

Regular sets are worked out from adjacent transactions in a stream which is known between a particular transaction in the past called a landmark and the existing transaction. While one steadily increases with the arrival of new transactions, the other endpoint (landmark) remains fixed. Transactions generally come in batches for this model that is generally appropriate for data streams.

h) Damped window model

Compared to the preceding windows the most new one have crucial importance in the damped windows. Older transactions do not add much towards the item set frequencies. For instance, the sliding window model will compute the average. The importance of data go slow exponentially into the past in the damped window model.

IV. Mining Frequent Item Set Over Data Streams: A Contemporary Affirmation of the State of Art

A regular item set mining algorithm over the data stream has been developed by Giannella et al [56]. Tilted windows have been used to estimate the regular patterns for the newest transactions built on the fact that the interest of the users lie on the newest transactions. Frequent item sets are symbolized by a tree data structure called as FP-stream, an increasing algorithm is used to maintain this. The earlier historical data is made use of to estimate the regular patterns increment by the implemented algorithm. A statistical technique was proposed by Laur et al [61] that is an unfair estimation of the support of sequential patterns or recall as selected by the user and restricts the degradation of the different principle. Using statistical support the conventional minimal support condition for checking regular sequential patterns was replaced. The theoretical foundations of the data stream analysis were checked by Gaber M M [49]. He also checked the mining data stream systems and techniques. The issues in the streaming was outlined and contemplated upon.

Observation: Algorithms [56] [13] [6] [49] that deal with mining frequent or sequential patterns on the data streams spotlight on how to technically deal with large historical data. There is no assurance that the algorithms can be run in limited memory capacity, when the data becomes huge and does not follow the running time.

MOMENT, an algorithm was suggested by Chi et al [63]. For continuous streaming data this showed sliding window to mine closed regular itemset. CET an internal data structure is used to check and store the closed item sets and boundary nodes. Teng [4] suggested a FTP-DS model that has sliding windows to ease the heavy information volume brought in by the continuous data streams. Data streams from the item sets give in temporal patterns to the FTP-DS mines. The sliding window data model is used by the FTP-DS an approximate mining algorithm. Li et al [11] created a DSM-FI technique to mine frequent item sets over the whole historical stream information. A distinctive top-down frequent item set discovery scheme and compact

pattern representation is accepted by the DSM-FI. Another algorithm on lossy counting was shown my Manku [57] and Motwani [13]. For the very first time mining, frequent item sets over the whole streaming data was used.

Observation: Historical Meta patterns got through data scanning are stored in internal data structures. This is a familiar feature that is explained in [4][11][63][13]. Additional procedures are created to re-mine the internal data structures when the user instigates a request for mining results.

Using a data structure called FP-Stream to keep historical knowledge such as item set frequencies Giannella et al [53] proved an approximation algorithm in random time intervals for mining frequent item sets. Linear Potential Function (LPF) algorithm to estimate propagation in Bayesian networks was shown by Zhang and Zhang [32]. An algorithm for keeping up frequent item sets in stream data in the supposition that every transaction holds a weight that is related to its age, permitting transactions on various ages to be dealt with different approaches was developed by Chang and Lee [38]. Stream an online algorithm was given by Asai et al [65] targeted mining patterns from semi-structured stream data like the XML data.

Observation: [53] [32] [38] [65] are the methodologies that permit handling of huge or infinite data streams in an estimated manner.

Cheung et al [42] [2] showed algorithms FUP and FUP2 for augmenting the frequent item sets. Similar algorithm was proposed by Thomas et al [12]. The mentioned models benefit from the tie up between the original database (DB) and incrementally changed transactions (db) and they also assume batch updates. Also known as Apriori algorithm FUP is a multiple-step algorithm. ZIGZAG was an algorithm proposed by Veloso et al [52] for mining regular item sets in developing databases. ZIGZAG was later extended by Otey et al [67] as parallel and distributed algorithms. [47] [2] and [12] have a striking similarity with Zigzag when both accelerate by using DB and db.

Observation: The main thing noted in the FUP is that some regular item sets will remain regular and some earlier irregular item sets will become regular when db is added to DB. Such item sets are called as winners. In addition some earlier regular item sets will become irregular. Such item sets are called as losers. The main method of FUP is to use the data in db to separate some winners and losers and thereby decrease the size of the candidate in Apriori algorithm. The working of the Apriori algorithm heavily relies on the size of the candidate set. Apriori's working is improved *to a large extent* by the FUP. Previous transactions from the database can be removed by extending FUP to FUP2. Hence FUP2 is used in the sliding-window data model and FUP will deal only with accumulative data model. FUP2 and the algorithm told in [12] are alike except that supplementing the regular item sets a negative border is retained. The regular item sets of db are mined first in the algorithm. Also the regular item sets in DB are updated in parallel. The regular item sets in the restructured database are calculated with a possible scan of the restructured database which is based on the change of the regular item sets in DB, the negative border in DB and the regular item sets in db. The algorithm [12] works well as the changed database is scanned at most once. Both accumulative and sliding window can make use of algorithm [12]. The [47], [2] and [12] come under the exact mining algorithm. ZIGZAG comes with many dissimilar features. Zigzag increases the speed of the support counting of common item sets in the reconstructed database and it does not find the regular item set in db itself. Hence with the lowest support ZIGZAG can take care of the batch update with random block size. It also takes in the techniques given by the GENMAX algorithm [31] and in every update it sustains maximum regular item sets. Sometimes this is not enough, at that point of time a second step is used in ZIGZAG. Here the restructured database is checked to find all regular item set and their supports.

An algorithm was created by Chi et al [16] where closed regular item sets are mined over data stream sliding windows. A condensed data structure called a closed enumeration tree (CET) was brought in to keep a vigorous selected set of item sets over the sliding window. There is a boundary in the closet regular item sets and the other item sets. The boundary movements in the CET will show the concept drifts in data streams. Sliding – window data model are the exact mining algorithms used by both the ZIGZAG and Moment.

1-pass algorithm, Count Sketch, was shown by Charikar et al [37] that resend the most regular items whose frequencies to convince an entry with high probabilities. A randomized algorithm was developed Manku et al [13] for upholding regular items over a data stream for a time t, the regular items defined over the whole data stream up to t.

Observation: No false negativity is promised by these algorithms and restriction the error of the calculated frequency. The Lossy Counting Algorithm is further improved to handle regular item sets. A process called as trie takes care of all regular item sets, this trie is restructured by batches of communications in the data stream. The algorithm Manku et al is attempt for a tunable negotiation between error bounds and memory usage. Accumulative data model is used by the Lossy Counting, Sticky Sampling and Count Sketch.

An algorithm presented by Chang et al [38] estDec, here frequency is defined by the age of a function. During a random time intervals, data streams mined by regular item sets, Giannella et al [56] proposed this algorithm. FP stream is helpful in storing

and updating historic data for the regular item sets and their occurrence with time and a period function is made use of to update the entries so that the newest entries are prejudiced more.

Observation: estDec and Giannella's algorithms are very prejudiced and are basically approximate mining algorithms. An error level in Giannella's algorithm is wee bit different as it gives error levels for information at different levels.

The numbers of patterns available have been vast. Showing the pattern in a condensed form is being worked upon. For example, Pei et al [19] and Zaki et al [18] show well-organized Closet and Charm. An item set is known as closed if all its supersets do not get the support that it gets [30]. Fp-tree was another efficient information structure that was coined by Han et al [57] for efficiently saving transactions for a specified low support threshold. A perfect algorithm known as FP-growth for mining Fp-tree was suggested.

Observation: Mining data streams does not get much support from the Fp-growth algorithm. In this case it goes through each window and is prohibitively expensive for big windows.

Windows methodology has gained a considerable amount of interest from regular item sets in data streams [43] [15] [7-9] [17] [59]. Therefore false negative based approach for mining of regular item sets over data streams has been proposed by Yu et al [1]. Lee at al [73] suggested the generation of k candidate sets from (k-1) candidate sets without checking their frequency. Jiang et al [17] recommended an algorithm for increasing holding the closed regular item sets over data streams. Extra passes over the data can be avoided this way and also this will not conclude in too many added candidate sets. Chi et al [63] advised the Moment algorithm for having a closed regular item sets over the sliding windows. When it comes to bigger slide sizes, Moment does not really help. Increasing mining on regular item sets is supported by Cats Tree [43] and Can Tree [15]. A good amount of work has been done by counting the candidate item sets more capably.

Observation: Park et al [8] suggested the hash-based counting method which was made use by several before mentioned regular item sets algorithms [7-2, 18, 8]. On the contrary Brin et al [27] suggested an active algorithm known as DIC for competently counting item sets frequencies. Fp-tree data structures are made use of by the fast verifiers in this model. The thought of putting conditions is to get must faster delta maintenance and counting patterns.

An association rule generation and summarization is another important area of work. Won et al [66] suggested a system for a hierarchical rule generation driven by ontology. In addition, Kumar et al [29] also suggested a single pass algorithm centered on the hierarchy-aware counting and transaction pruning for mining association rules, after the structure is given. Liu et al [64] recommended a methodology to systematize and sum up the found association rules. This methodology simplifies the rules and keeps stays on exceptions to the generalization.

Observation: [66] focuses on holding the level of items and the categorization of rules making use of a hierarchical association rule collecting the groups the created rules from the item space to the hierarchical space. To find comprehensive association, simplified methods [29] can be used over the history of organization. Based on appealing measures [21] [71] many research projects in this arena have looked upon the rule ordering. For example, Li et al [21] suggested rule position depending on assurance, sustenance and the number of items on the left hand side to trim the found rules.

In the past 10 years, a good number of algorithms have been suggested on the mining regular patterns in data streams. These can be divided into three special categories, landmark based [4, 13, 1, 9] damped or time decay based [39, 33] and a sliding window based [26-34] algorithms. DSM-FI [11] is an innovative algorithm which converts all operations into minor transactions and is put in the synopsis data structure known as the item-suffix regular item set forest that is based on the prefix-tree. In order to get fairly accurate results of regular patterns over the landmark window the authors made use of Chernoff Bound [1]. Lattice structure was made use by Zhi-Jun et al this is also known as the regular enumerate tree that is separated into many equal classes of the accumulated patterns with the similar transaction ids in a single class. estDec was recommended by Change and Lee reproduced the time decay model where every transaction has an influence that lessens with age [39]. Algorithm [33] which is alike estDec, it is said that mining increasing number of regular item set instantaneously over data stream based on a damped model.

Observation: The data that enters from a particular time called landmark till the existing time is taken into consideration in the landmark model. This time can be the starting or the restarting time. Earlier and existing transactions of the input stream are considered similar. Regular models are further separated into equivalent classes and only these regular patterns symbolize the two borders of every class are preserved, other regular models are trimmed.

Significance is given to data elements based on their arrival order in the time decal model. This is done in order to highlight the recently arrived data. A decay rate is shown to lessen the effect of the previous transaction in the set of regular pattern.

Numerous sliding window based algorithms recommended for regular item set mining over data streams. Raw transactions of sliding windows can be stored using DS Tree [26] and CPS-Tree [74]. Fixed tree

structure is used by DS Tree in canonical order of the branches and CPS-Tree is recreated to keep a check on the memory usage. Both these use the FP-Growth [418] for mining as suggested for static databases. MFI-Trans based on priori algorithm was recommended in [54]. Every regular item sets is mined over the newest window of transactions. Bit string is made use of for all items to keep its happening data in the window. A sliding based algorithm has been recommended in [28]. This has a window content that is kept vigorously using a set of easy lists.

Lin et al [51] suggested a fresh technique for mining regular patterns on time receptive sliding window. Here window is divided into batches and mining for this item set is done discretely. A limit in frequent closed item set and other item set is kept; this helps the Moment algorithm to locate the closed regular item sets.

The regular item set in one pane of the window are taken into consideration for further check to look out for regular item sets in the entire window in SWIM [77] a pane based algorithm. The union of these regular patterns of all panes is maintained. It also increasing updates support and trims irregular ones. Transactions are stored as a prefix tree for each plane. The authors devised algorithm for mining [4-10] constantly to maintain the non-derivable regular item sets of the sliding windows. Closed regular item sets and non-derivable can be viewed as a synopsis of all regular item sets.

estWin algorithm was recommended by Chang and Lee it locates the newest regular patterns adaptively on transactional data streams making use of the sliding window model. Monitoring lattice is made use of as a prefix tree to check the set of regular item sets over a data stream. Every node of the lattice will symbolize an item set that can be created by using items kept in the nodes in the way from the root to the node. All the data is kept in the final node of the item set. A Less amount and small support called as minimum important is used by the algorithm to know the new regular item set to approximate their support. The existing set of regular item set is updates going to the related item set in the monitoring lattice, when a novel transaction come from the input data stream. With this noteworthy item sets are known by second traversal of the related ways of the monitoring lattice making use of a transaction. In order to remove their effect when the algorithm expires keeps the set of transaction of the window in the memory. Connected ways of the monitoring lattice must be gone through to delete the previous transaction from the window. The algorithm trims irrelevant item sets from the tree after reaching a stable number of transactions making use of a method called as force pruning. When all the subsets become irrelevant and are kept in the tree an item set is put to the prefix tree. The support of the novel important item set in the old transaction of the window is estimated by the algorithm. The item set of length k (k-item set) calculated support is alike to the lowest support in all its subsets with length k-1. Information like possible count 9pcnt), error (err), actual count (acnt) and first transaction (mid) are available in each node of the prefix tree. Pcnt is known as the calculated frequency of the item set in the older transaction of the window before the item set was put to the prefix tree. Acnt is the checked frequency of the item set post its insertion in the tree. Err is termed as the error of the calculated support of the item set estimated using the subsets. Id of transaction that results the item set to be put in the tree is known as Mtid. The total of Acnt and Pcnt gives support of an item set in the prefix tree. To know more about data computing the probable count and the error in support of important item sets and other features of the estWin algorithm we can refer to [34]. On putting the item set, fresh transactions come in, previous transactions are removed and the error is minimized. Due to the removal of previous transactions the error will lessen.

New-Moment was used by H F Li [45] to preserve a set of regular closed item sets in data streams with transaction sensitive sliding window. Chi recommended MOMENT [63] which was a usual algorithm that can reduce the size of the data structure. A new way was suggested by N Jiang [70] for mining regular closed item sets over data streams. Compact data structure was brought in by Y Chi [16], this is a closed enumeration tree that keep a set of item sets that are dynamically selected in a sliding window. There is a border between the item sets and regularly closed item sets. FPCFI-DS was an algorithm suggested by F J Ao [10] for mining closed regular item sets in data streams. This makes use of single-pass lexicon graphical order FP-Tree-based algorithm that is merged with the ordering policy to mine the closed regular item sets in the primary window and the tree is updated for every sliding window.

Another mining task for recommended by J Y wang [T8-9] for mining top-k regular closed item sets whose length was greater than min_l.

Observation: compared to the sliding window based regular item set the data streams can be categorized as two groups. [26] [74] [54] and [28] is the first group. The window is kept above the newest transactions and frequent patterns in the current window are extracted by the mining process. This happens when the user submits a request. Hence the intention here is to keep the transactions making use of the low memory and do the mining process competently. [51] [9-8] [77] [41] and [34] is the second group where the mining results are constantly updated by making use of the fresh transaction to the window and previous transactions are removed from the window. Hence quick updation of the mining result and less memory usage gives the desired result. As you can see the mining

result with immediate effect the algorithms of the second group are proven to be more useful. To give out quickly the estimated algorithm that can find set of regular item sets in the sliding window with elevated quality of result is adequate. The first group [26-6] fails to adaptively maintain and update the mining result. The result becomes invalid when fresh transactions arrive from the stream. This results in the re-execution of the mining task. Conversely, relating a mining algorithm to the entire window will require significant processing and memory requirements specifically in big window size in respect to algorithms of second group [28] [51] [9-8] [77] and [410] where mining results are updated adaptively. Compared to the high arrival rated of the input streams the functioning of these algorithms are poor.

In [69], [75] and [44] sequential pattern-mining algorithms in data streams are presented. The issues relating to the mining distributed data stream was dealt in many areas of data mining such as mining regular item sets [55], [5], clustering [24],[3] and association rules [7], [50]. In contrast, sequential patterns of mining in several data streams are dealt in [72] and [42]. Collective Data Mining (CDM) has been suggested in [22].

Observation: The methods [69], [75] and [44] cannot deal with the distributed streams, as accumulating data streams in one location before giving them out come with a high-priced cost and can invade owner's privacy. Old mining results will not be preserved in MILE algorithm, which happens to be one-time fashioned algorithm. Therefore, this might take excessive amount of time in re-mining. In IAs pam algorithm this issue has been done away with. It was shown in [42] that increasingly mine across-streams sequential type to hold the fresh mining results. Nevertheless centralized setting strategy is made use here and it depends on a sliding window at the same working node to read the sample data streams. CDM has a purpose to guarantee that incomplete models made from local data at various sites are right. A global model can be formed from these models.

V. Conclusion

The deliberations above we observe that the present mining strategies have an increasing and one pass mining algorithm that are appropriate to mine data streams. Very few of them tackle the concept drifting problem. Fairly accurate results are obtained from these algorithms as a result of large data streams and less memory. Unlike traditional databases, here we cannot keep the frequency counts of all item sets in the entire data streams. Precise mining results are obtained by some suggested algorithms by keeping up to a minute subset of regular item sets from data streams, retaining their precise frequency counts. Sliding window data processing model helps in preserves only some section of the regular item sets in the sliding windows. Different other ways to maintain is maximal frequent item sets, closed frequent item sets and short frequent item sets.

The existing stream data mining techniques need users to describe one or more parameters before its implementation. But many of these streams do not tell how these parameters can be adjusted while they run. Waiting for the mining algorithm to cease to reset the parameters is not a good idea as it will take a long time for the execution as the data is massive.

Some recommended techniques permit the user to regulate some parameters online. There is no assurance that these parameters are of key importance and they might not be helpful in the mining scenario. Auto adjustment of mining algorithms and users adjusting online was also considered. The research in this field is in gestation stage. To tackle all those discussed issues in this paper would propel the process of developing association rule mining applications in data stream systems. Previous researches focused on creating competent mining techniques. A new aspect of data mining streams is necessary to find scalable regular item set mining models with the following factors.

1. Utility Mining
2. Weighted Supports
3. Multiple Supports
4. Transitional States
5. Temporal Validity

If many of these issues are yet too solved and a perfect and efficient system has to be developed, it is certain that in coming future data stream item set will be very important in the business world. Focusing on this we guesstimate more research in this direction.

References Références Referencias

1. Information Sciences, 176(14), 2006, pp. 1986–2015.
2. A General Incremental unquie for Maintaining Discovered Association Rules. Proceedings of the Fifth International Conference on Database Systems for Advanced Ap- plications (DASFAA), Melbourne, Australia, April 1, 4, 1997.
3. A Generic Local Algorithm for Mining Data Streams in Large Distributed Systems
4. A regression-based temporal pattern mining scheme for data streams. In: Proceedings of the 29th VLDB Conference (Berlin, Germany, 2003) 93–104.
5. A Sketch-Based Architecture for Mining Frequent Items and Item sets from Distributed Data Streams. In: Proc. of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (2011), pp. 245-253.

6. Alternative Interest Measures for Mining Associations. In IEEE TKDE, 15:57-69, 2003.
7. An Approach for Distributed Streams Mining Using Combination of Naïve Bayes and Decision Trees
8. An effective hash-based algorithm for mining association rules. In SIGMOD, pages 175-186, 1995.
9. An efficient algorithm for frequent itemset mining on data streams, Proc. ICDM, 2006, pp. 474–491.
10. itemsets in data Streams. Proceedings of the IEEE 8th International Conference on Computer and Information Technology, pp.37-42, 2008.
11. An efficient algorithm for mining frequent itemsets over the entire history of data streams. In: 1st International Workshop on Knowledge Discovery in Data Streams (Pisa, Italy, 2004) 20–24.
12. An Efficient Algorithm for the Incremental Updation of Association Rules in Large Databases. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), Newport Beach, California, USA, August 14, 17, 1997.
13. Approximate frequency counts over data streams. In: Proceedings of the 28th VLDB Conference (Hong Kong, China, 2002).
14. Beyond Market Basket: Generalizing Association Rules to Correlations. In Proc. of SIGMOD, 1997.
15. Cantree: A tree structure for efficient incremental mining of frequent patterns. In ICDM, pages 274-281, 2005.
16. Catch the Moment: Maintaining Closed Frequent Itemsets in a Data Stream Sliding Win- dow. Knowledge and Information Systems, 10(3): 265-294.
17. Cfi-stream: mining closed frequent itemsets in data streams. In SIGKDD, pages 592-597, 2006.
18. CHARM: An efficient algorithm for closed itemset mining. In SIAM, 2002.
19. CLOSET: An efficient algorithm for mining frequent closed itemsets. In SIGMOD, pages 21-30, 2000.
20. Clustering by Pattern Similarity in Large Datasets. In Proc. of SIGMOD, 2002.
21. CMAR: Accurate and efficient classification based on multiple class-association rules. In ICDM, pages 369-376, 2001.
22. Collective Data Mining: A New Perspective towards Distributed Data Mining. In: Advances in Distributed and Parallel Knowledge Discovery, editied by H. Kargupta and P. Chan, chapter, 5, AAAI Press / The MIT Press (2000).
23. Data Streams and Histograms; ACM Symposium on Theory of Computing; 2001.
24. Discovering Trend-Based Clusters in Spatially Distributed Data Streams. International Workshop of Mining Ubiquitous and Social Environments (2010), pp 107-122.
25. Discovery of Frequent Episodes in Event Sequences. In DMKD, 1:259-289, 1997.
26. DSTree: a tree structure for the mining of frequent sets from data streams", in Proceedings of IEEE International Conference on Data Mining, 2006, pp. 928–932.
27. Dynamic itemset counting and implication rules for market basket data. In SIGMOD, pages 255-264, 1997.
28. EclatDS: An Efficient Sliding Window Based Frequent Pattern Mining Method for Data Streams", Intelligent Data Analysis, Vol. 15(4), 2011, pp. 571-587.
29. Efficient algorithm for hierarchical online mining of association rules
30. Efficient mining of association rules using closed itemset lattices. Information Systems, pages 25-46, 1999.
31. Efficiently Mining Maximal Frequent Itemsets. Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, California, USA, November 29 , December 2, 2001.
32. Encoding probability propagation in belief networks IEEE Transactions on Systems, Man and Cybernetics (Part A) 32(4) (2002) 526–31.
33. est Max: Tracing Maximal Frequent Itemsets Instantly over Online Transactional Data Streams", IEEE Transactions on Knowledge and Data Engineering, 21 (10), 2009, pp. 1418-1431.
34. estWin: Online data stream mining of recent frequent itemsets by sliding window method", Journal of Information Science, Vol. 31, 2005, pp. 76–90.
35. Fast Algorithms for Mining Association Rules.
36. Fast algorithms for mining association rules" in Proceedings of International Conference on Very Large Databases, 1994, pp. 487–499.
37. Finding Frequent Items in Data Streams. Proceedings of the 2002 International Collo- quium on Automata, Languages and Programming, Malaga, Spain, July 8, 13, 2002.
38. Finding recent frequent itemsets adaptively over online data streams. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2003) (Washington, DC, 2003) 226–35.
39. Finding recently frequent itemsets adaptively over online transactional data streams Frequent Itemset Mining Implementations. In Proc. of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, 2003.
40. Frequent pattern mining: current status and future directions", Data Mining and Knowledge Discovery, Vol. 15, 2007, pp. 55–86.
41. Incremental Mining of Across-streams Sequential Patterns in Multiple Data Streams. Journal of Computers Vol. 6 (2011), pp.449-457.

42. Incremental mining of frequent patterns without candidate generation or support. In 7th Database Engineering and Applications Symposium International Proceedings, pages 111-116, 2003.
43. Incremental Mining of Sequential Patterns over a Stream Sliding Window. In: Proc. of the 6th IEEE International Conference on Data Mining (2006), pp.677-681.
44. Incremental updates of closed frequent itemsets over continuous data streams. Expert Systems with Applications, Vol.36, pp.2451-2458, 2009.
45. Integrating Classification and Association Rule Mining. In Proc. of KDD, 1998.
46. Maintenance of Discovered Association Rules in Large Databases: An Incremental Up - dating Technique. Proceedings of the Twelfth International Conference on Data Engineering, New Orleans, Louisiana, USA, February 26 , March 1, 1996.
47. Mining Association Rules between Sets of Items in Large Databases. In Proc. of SIGMOD, 1993.
48. Mining Data Streams: A Review, Mining Distributed Evolving Data Streams using Fractal GP Ensembles. In: Proc. of the 10th European Conference on Genetic Programming (2007), pp. 160-169.
49. Mining frequent itemsets from data streams with a time-sensitive sliding window", in Proceedings of SDM International Conference on Data Mining, 2005.
50. Mining Frequent Itemsets in Evolving Databases. Proceedings of the Second SIAM International Conference on Data Mining, Arlington, VA, USA, April 11, 13, 2002.
51. Mining frequent itemsets over arbitrary time intervals in data streams. In: Technical Report TR587 (Indiana University, IN 2003).
52. Mining frequent itemsets over data streams using efficient window sliding techniques", Expert Systems with Applications, Vol. 36, 2009, pp. 1466–1477.
53. Mining Frequent Itemsets over Distributed Data Streams by Continuously Maintaining a Global Synopsis. Data Mining and Knowledge Discovery Vol. 23 (2011), pp. 252-299.
54. Mining Frequent Patterns in Data Streams at Multiple Time Granularities,
55. Mining frequent patterns without candidate generation. In: SIGMOD' 00 (ACM Press, Dallas, TX, 2000) 1–12.
56. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, Vol. 8, 2004, pp. 53-87.
57. Mining maximal frequent itemsets from data streams. Information Science, pages 251-262, 2007
58. Mining non-derivable frequent itemsets over data stream", Data & Knowledge Engineering, Vol. 68, 2009, pp. 481-498.
59. Mining Sequential Patterns on Data Streams: a Near-Optimal Statistical Approach.
60. Mining Sequential Patterns. In Proc. of IDCE, 1995.
61. Moment: maintaining closed frequent itemsets over a stream sliding window. In: Rajeev Rastogi et al. (eds) Proceedings of 4th IEEE International Conference on Data Mining (Brighton, UK, 2004) 59–66.
62. Multi-level organization and summarization of the discovered rules. In Knowledge Discovery and Data Mining, pages 208-217, 2000.
63. Online algorithms for mining semi structured data streams.
64. Ontology-driven rule generalization and categorization for market data. Data Engineering Workshop, 2007 IEEE 23rd International Conference on, pages 917-923, 2007.
65. Parallel and Distributed Methods for Incremental Frequent Itemset Mining. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 34(6): 2439, 2450.
66. Querying and Mining Data Streams: You Only Get One Look. In Tutorial of SIGMOD, 2002.
67. Random Sampling Over Data Streams for Sequential Pattern Mining. In: Proc. of the 1st European Workshop on Data Streams (2007), pp. 61-66.
68. Research issues in data stream association rule mining. SIGMOD Record 35 (1), pp.14-19, 2006.
69. Selecting the right interestingness measure for association patterns. In Knowledge Discovery and Data Mining, pages 32-41, 2002.
70. Sequential Pattern Mining in Multiple Streams. In: Proc. of the 5th IEEE International Conference on Data Mining (2005).
71. Sliding window filtering: an efficient method for incremental mining on a time-variant database. Information Systems, pages 227-244, 2005.
72. Sliding window-based frequent pattern mining over data streams", Information Sciences, Vol. 179, 2009, pp. 3843-3865.
73. Stream Sequential Pattern Mining with Precise Error Bounds. In: Proc. of the IEEE International Conference on Data Mining (2008), pp. 941-946.
74. The space complexity of approximating the frequency moments. Compute. Syst. Sci., 58(1):137–147, 1999.
75. Verifying and mining frequent patterns from large windows over data streams", in Proceedings of International Conference on Data Engineering, 2008, pp. 179–188.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 5 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Survey on Enhancing Cloud Data Security using EAP with Rijndael Encryption Algorithm

By Sanjoli Singla & Jasmeet Singh
RIMT-IET, Mandi Gobindgarh, India

Abstract - Cloud computing is an amazing technology where users can remotely store their data into the cloud to enjoy high quality application and services. Cloud being the most vulnerable next generation architecture consist of two major design elements i.e. the Cloud Service Provider(CSP) and the Client. Even though the cloud computing is promising and efficient, there are many challenges for data privacy and security. This paper explores the security of data at rest as well as security of data while moving.

Keywords : *authentication, cloud, encryption, rijndael algorithm.*

GJCST-C Classification *: H.2.7*

Strictly as per the compliance and regulations of:

Survey on Enhancing Cloud Data Security using EAP with Rijndael Encryption Algorithm

Sanjoli Singla [α] & Jasmeet Singh [σ]

Abstract - **Cloud computing is an amazing technology where users can remotely store their data into the cloud to enjoy high quality application and services. Cloud being the most vulnerable next generation architecture consist of two major design elements i.e. the Cloud Service Provider(CSP) and the Client. Even though the cloud computing is promising and efficient, there are many challenges for data privacy and security. This paper explores the security of data at rest as well as security of data while moving.**

Keywords : authentication, cloud, encryption, rijndael algorithm.

I. Introduction

The cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable resources such as networks, servers, files storage, applications and services. The cloud computing field is growing day by day with a increasing number of businesses and government establishments going for cloud computing based services.[1]The cloud computing incorporates combination of:-

1. IaaS (Infrastructure as a Service)
2. PaaS (Platform as a Service)
3. SaaS (Software as a Service)

These are collectively called as *aaS (Everything as a Service) which means a service-oriented architecture. The first benefit of the cloud computing is that it reduces the cost of hardware used at user end. As the data is stored at the cloud, so instead of buying the whole infrastructure required to run the processes user just rent the assets according to his requirement. The similar idea is behind all cloud networks as shown in figure 1. [2]

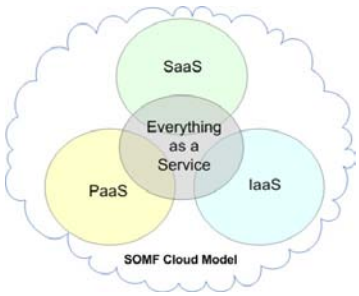


Figure 1 : Cloud Service Models

Author α : M.TECH (Computer Science and Engineering) RIMT-IET, Mandi Gobindgarh, Punjab India. E-mail : sanjoli_11@yahoo.co.in
Author σ : Asst. Professor (CSE Department) RIMT-IET, Mandi Gobindgarh, Punjab India. E-mail : jasmeetgurm@gmail.com

Cloud Computing deployment models are:-

Private Cloud

Enterprise owner or leased.

Community Cloud

Cloud infrastructure that supports a specific community with shared concerns.

Public Cloud

Available to the public and owned by an organization selling cloud services.

Hybrid Cloud

Composition of two or more clouds(Private, Community, Public). [4]

a) Data security issues in the cloud

Securing data is always of vital importance and because of the critical nature of cloud computing and large amounts of complex data it carries, the need is even important. Therefore, data privacy and security are issues that need to be resolved as they are acting as a major obstacle in the adoption of cloud computing services. The major security issues with cloud are:-

i. *Privacy and Confidentiality*

Once the clients outsource data to the cloud there must be some assurance that data is accessible to only authorized users. The cloud user should be assured that data stored on the cloud will be confidential.

ii. *Security and Data Integrity*

Data security can be provided using various encryption and decryption techniques. With providing the security of the data, cloud service provider should also implement mechanism to monitor integrity of the data at the cloud. [3]

II. Related Work

In [3] paper author's main concern with reference to the security of the data is how to ensure security of data at rest. In it RSA (Ron Rivest, Adi Shamir and Len Adleman) algorithm is used which consist of public key and private key. Encryption is done by Cloud Service Provider using public key and decryption is done by cloud using corresponding private key.

In [8] paper author main focus is on service provider side security. To ensure security (Privacy & Confidentiality) he proposes EAP with Challenge Handshaking Protocol for the authorization of the user and RSA for encryption at server side.

In [1] paper author proposes the combination of Rijndael with Digital Signature for enhancement of security. CSP first converts the data into hash and then encrypt it using digital signature and finally encrypt using Rijndael algorithm with user's public key. On the other side user can decrypt data with its private key and uses CSP's public key for verification of signature.

In all the approaches mentioned above we analyze that security is only provided to data at rest i.e. encryption is done at the cloud side. So when the user outsources the data to the cloud, data can be attacked while on the way. To resolve this issue encryption must be done at the user side. So that data can travel in encrypted form only. For this strong encryption algorithm i.e. Rijndael Encryption Algorithm can be used. Also to prevent user's data from unauthorized access EAP-CHAP can be used.

III. Techniques

a) Authentication Protocol

EAP will implement on Cloud environment for authentication purpose. However different categories EAP are classified by authentication method. In our purposed model we use Challenge-Handshake Authentication Protocol (CHAP) for authentication. When client demands data or any service of cloud computing. Service Provider Authenticator (SPA) first requests for client identity. The whole process between client and Cloud provide explain in a figure 2 given below.

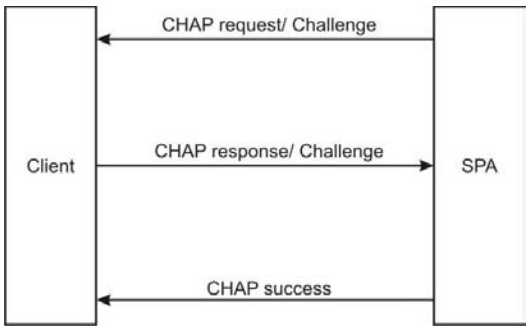


Figure 2 : Implementation of CHAP in cloud computing

Authentication of CHAP performs in three steps

1. When client demands a service, Service Provider Authentication sends a "challenge" message to client.
2. Client responds with a value that is calculated by using one way hash function on the challenge.
3. Authenticator verifies the response value against its own calculated hash value. If the values match, the Cloud provider will give service, otherwise it should terminate the connection.

Implementation of EAP-CHAP in Cloud Computing will solve the authentication and authorization problems. [8]

b) Rijndael Encryption Algorithm

Rijndael is the block cipher algorithm recently chosen by the National Institute of Science and Technology (NIST) as the Advanced Encryption Standard (AES). It supersedes the Data Encryption Standard (DES). Rijndael is a standard symmetric key encryption algorithm used to encrypt sensitive information. The choice was based on a careful and comprehensive analysis of the security and efficiency characteristics of Rijndael's algorithm.

Rijndael is an iterated block cipher. Therefore, the encryption or decryption of a block of data is accomplished by the iteration (a round) of a specific transformation (a round function). Rijndael also defines a method to generate a series of sub keys from the original key. The generated sub keys are used as input with the round function. Rijndael was designed based on the following three criteria:

1. Resistance against all known attacks;
2. Speed and code compactness on a wide range of platforms;
3. Design simplicity

Rijndael is the best combination of security, performance, efficiency, ease of implementation and flexibility. The Rijndael algorithm supports key sizes of 128, 192 and 256 bits, with data handled in 128-bit blocks. Rijndael uses a variable number of rounds, depending on key/block sizes, as follows: 9 rounds if the key/block size is 128 bits, 11 rounds if the key/block size is 192 bits, and 13 rounds if the key/block size is 256 bits.

Rijndael is a substitution linear transformation cipher. It uses triple discreet invertible uniform transformations (layers). Specifically, these are: Linear Mix Transform; Non-linear Transform and Key Addition Transform. Even before the first round, a simple key addition layer is performed, which adds to security. Thereafter, there are Nr-1 rounds and then the final round. The transformations form a State when started but before completion of the entire process. [1]

i. *High-level description of the algorithm*

a. *Key Expansion*

Round keys are derived from the cipher key using Rijndael's Key schedule.

b. *Initial Round*

Add Round Key each byte of the state is combined with the round key using bitwise xor.

c. *Rounds*

- SubBytes
- ShiftRows
- MixColumns
- AddRoundKey

d. *Final Round (no MixColumns)*

- SubBytes
- ShiftRows

- Add Round Key

e. *The SubBytes Step*

The SubByte step is a non-linear byte substitution that operates on each of the 'state' bytes independently, where a state is an intermediate cipher result. Here each byte in the state matrix is replaced with a SubByte using an 8-bit substitution box, the Rijndael S-box. The S-box used is derived from the multiplicative inverse over GF(2^8) and then an affine transformation is applied.

f. *The ShiftRows Step*

The ShiftRows step operates on the rows of the state; it cyclically shifts the bytes in each row by a certain offset. For AES, the first row is left unchanged. Each byte of the second row is shifted one to the left. Similarly, the third and fourth rows are shifted by offsets of two and three respectively. For blocks of sizes 128 bits and 192 bits, the shifting pattern is the same. Row n is shifted left circular by n-1 bytes. In this way, each column of the output state of the ShiftRows step is composed of bytes from each column of the input state. Rijndael variants with a larger block size have slightly different offsets. For a 256-bit block, the first row is unchanged and the shifting for the second, third and fourth row is 1 byte, 3 bytes and 4 bytes respectively—this change only applies for the Rijndael cipher when used with a 256-bit block as AES does not use 256-bit blocks.

g. *The MixColumns Step*

In the MixColumns step, the four bytes of each column of the state are combined using an invertible linear tranformation. The MixColumns function takes four bytes as input and outputs four bytes, where each input byte affects all four output bytes. Together with ShiftRows, MixColumns provides diffusion in the cipher.

During this operation, each column is multiplied by the known matrix that for the 128-bit key is:

$$\begin{bmatrix} 2 & 3 & 1 & 1 \\ 1 & 2 & 3 & 1 \\ 1 & 1 & 2 & 3 \\ 3 & 1 & 1 & 2 \end{bmatrix}.$$

The multiplication operation is defined as: multiplication by 1 means no change, multiplication by 2 means shifting to the left, and multiplication by 3 means shifting to the left and then performing xor with the initial unshifted value. After shifting, a conditional xor with 0x11B should be performed if the shifted value is larger than 0xFF.

In more general sense, each column is treated as a polynomial over GF (2^8) and is then multiplied modulo x^4+1 with a fixed polynomial $c(x) = 0x03 \cdot x^3 + x^2 + x + 0x02$. The coefficients are displayed in their hexadecimal equivalent of the binary representation of bit polynomials from GF(2)[x]. The MixColumns step can also be viewed as a multiplication by a particular MDS matrix in a finite field.

h. *The AddRoundKey step*

In the AddRoundKey step, the subkey is combined with the state. For each round, a subkey is derived from the main key using Rijndael's key schedule; each subkey is the same size as the state. The subkey is added by combining each byte of the state with the corresponding byte of the subkey using bitwise XOR. [11]

IV. Conclusion

Data security has become the vital issue of cloud computing security. It depends upon the way Cloud Service Provider (CSP) allows its client to get registered with his cloud network. In our survey we analyze how security is provided to the data at rest i.e. encryption is done by the cloud service provider. But we noticed that this is not sufficient as when the user outsources the data to the cloud data must also be secured on the way to the cloud. So it is preferred that encryption must be done by the user to provide better security.

References Références Referencias

1. Prashant Rewagad, Yogita Pawar, "Use of Digital Signature and Rijndael encryption Algorithm to Enhanced Security of data in Cloud computing Services", proceeding published in International Journal of Computer Applications (IJCA), 2012.
2. Mandeep Kaur, Manish Mahajan, "Using encryption Algorithms to enhance the Data security in Cloud computing", International Journal of Communication and Computer Technologies (IJCCTS), Vol. 01, Issue: 03, January 2013.
3. Parsi Kalpana, Sudha Singaraju, "Data Security in Cloud Computing using RSA Algorithm", International Journal of Research in Computer and Communication Technology (IJRCCT) , Vol. 1, Issue 4, September 2012.
4. Eman M.Mohamed, Hatem S.Abdelkader,Sherif El-Etriby, "Data Security Model for Cloud Computing", The Twelfth International Conference on Networks: ICN 2013.
5. C. Bagyalakshmi, Dr. R. Manicka Chezian, "A Survey on Cloud Data Security using Encryption Technique", International Journal of Advanced Research in Computer Engineering and Technology, Vol. 1, Issue 5, July 2012.
6. Jiyi Wu, Qianli Shen, Tong Wang, Ji Zhu, Jianlin Zhang, "Recent Advances in Cloud Security" Journal of Computers, Vol. 6, No. 10, October 2011.

7. Muneshwara M.S, Arvind Tejas Chandral, "Monitoring the integrity of dynamic data stored in Cloud Computing", International Journal of Engineering Research & Technology (IJERT), Vol. 1, Issue 4, June 2012.
8. Sadia Marium, Qamar Nazir, Aftab Ahmed, Saira Ahthasham, Mirza Aamir Mehmood," Implementation of EAP with RSA for enhancing the security of cloud computing", International Journal of Basics and Applied Sciences, 1 (3) 2012 177-183.
9. http://www.efgh.com/software/rijndael.htm
10. http://en.wikipedia.org/wiki/Cloud_computing
11. http://en.wikipedia.org/wiki/Advanced_Encryption_Standard

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 5 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

ECG Signal Denoising by Morphological Top-Hat Transform

By S.A.Taouli & F. Bereksi-Reguig
University Aboubekr Belkaid

Abstract - The electrocardiogram ECG signal plays an important role in the primary diagnosis, prognosis and survival analysis of heart diseases. The ECG signal contains an important amount of information that can be exploited in different manners. However, during its acquisition it is often contaminated with different sources of noise making difficult its interpretation. In this paper, a new approach based on Morphological Top-Hat Transform (MTHT) is developed in order to suppress noises from the ECG signals. The morphological operators (dilation, erosion, opening, closing) constitute the fundamental stage of Top-Hat transform. Method presented in this paper is compared with the Visu Shrink, Sure Shrink, and Bayes Shrink methods. The experimental results indicated that the proposed methods in this work were better than the compared methods in terms of retaining the geometrical characteristics of the ECG signal, SNR. Due to its simplicity and its fast implementation, the method can easily be used in clinical medicine.

Keywords : *ECG signal, denoising, mathematical morphology, morphological top-hat transform.*

GJCST-C Classification : H.5.5

Strictly as per the compliance and regulations of:

ECG Signal Denoising by Morphological Top-Hat Transform

S.A.Taouli [α] & F. Bereksi-Reguig [σ]

Abstract - The electrocardiogram ECG signal plays an important role in the primary diagnosis, prognosis and survival analysis of heart diseases. The ECG signal contains an important amount of information that can be exploited in different manners. However, during its acquisition it is often contaminated with different sources of noise making difficult its interpretation. In this paper, a new approach based on Morphological Top-Hat Transform (MTHT) is developed in order to suppress noises from the ECG signals. The morphological operators (dilation, erosion, opening, closing) constitute the fundamental stage of Top-Hat transform. Method presented in this paper is compared with the Visu Shrink, Sure Shrink, and Bayes Shrink methods. The experimental results indicated that the proposed methods in this work were better than the compared methods in terms of retaining the geometrical characteristics of the ECG signal, SNR. Due to its simplicity and its fast implementation, the method can easily be used in clinical medicine.

Keywords : ECG signal, denoising, mathematical morphology, morphological top-hat transform.

I. Introduction

During its acquisition the ECG signal is corrupted with different types of noises. Noises such as the power line interference (50 Hz), the muscle artifact due to the EMG (electromyogram), the baseline wandering due to the rythmic inhalation and exhalation during respiration are examples of noises which corrupt the ECG signals [1-2]. In order to reduce the noise in ECG signals many techniques are available such as digital filters (FIR or IIR), adaptive method, wavelet transform thresholding and Empirical Mode Decomposition methods [3]. However, digital filters and adaptive methods can be applied to signal whose statistical characteristics are stationary in many cases.

Recently the wavelet transform has been proven to be an useful tool for non-stationary signal analysis [4]. Thresholding is used in wavelet domain to smooth out or to remove some coefficients of wavelet transform subsignals of the measured signal. The noise content of the signal is reduced, effectively, with in the non-stationary environment. The denoising method that applies thresholding in wavelet domain has been proposed by Donoho [5-6]. It has been proved has Been proved that the Donoho's method for noise reduction works well for a wide class of one dimensional and two dimensional signals. Other approaches for threshold value estimators can be found in [7-10]. Visu Shrink [9, 11] utilizes the universal threshold estimator, which is $\sqrt{2\log(N)}$ for a vector d_i of the detail coefficients of length N . Sure Shrink is based on Stein's unbiased risk estimator [12]. Sure Shrink has serious drawbacks in situations of extreme sparsity of the wavelet coefficients [13]. In [11], Bayes Shrink was used for the threshold estimator, which is a data-driven sub and adaptive technique. Others methods, which has also been widely used is the Least Mean Square adaptive algorithm (LMS) [9]. But this algorithm is not able to track the rapidly varying non-stationary signals such as ECG signal within each heart beat; this causes excessive low pass filtering of mean parameters such as QRS complex.

This paper considers as a possible alternative, the application of Morphological Top-Hat Transform, (MTHT) based on morphological signal-processing concepts. Morphological signal processing comprises a broad collection of theoretical concepts and mathematical tools for signal analysis, non-linear signal operators, design methodologies and application systems that are related to Mathematical Morphology (MM).

Morphological operators of opening and closing are simple, and the morphological Top-Hat transform arising from these operators, have been prove to be powerful tools and have been used in different applications, giving excellent results in areas such as noise reduction, edge detection and object recognition [14].

In this work, we are interested in morphological Top-Hat transform. The filter is implemented under MATLAB 7 environment. The filter is evaluated using ECG signals from the MIT- BIH universal data base [15].

The paper is divided into five sections. After this introduction section 1, the Section 2 presents a brief description of basic morphological signal processing. Section 3 describes the morphological Top-Hat transform algorithm for ECG signal. In section 4, some experimental results are presented and discussed. Finally, a conclusion is given in section 5.

Author α : Teacher and Researcher of Biomedical Engineering Research Laboratory, Dept. of genius electric and electronic, Tlemcen-algeria. E-mail : s_taouli@mail.univ-Tlemcen.dz

Author σ : Prof. and director of Biomedical Engineering Research Laboratory. E-mail : fethi.bereksi@mail.univ-Tlemcen.dz

II. Mathematical Morphology

During the last decade, mathematical morphology was established like powerful method for signal processing and became a complete mathematical theory. It was applied successfully in various disciplines, such as mineralogy, medical diagnosis, and histology. It found also increasing applications in the digital signal processing, computer vision and pattern recognition.

a) Basic mathematical morphology transforms

The basic concept of morphological signal processing is to modify the shape of a signal, equivalently considered as a set, by transforming it through its interaction with another object, called structuring element. In practice, the structuring element is compact and of a simple shape than the original object. The basic operators of morphology transform include dilation, erosion, opening and closing [16-19].

Let $f(n)$ be the original 1-D signal, which is the discrete function over a domain $f(n)=\{0,1.....,N-1\}$. And let $B(m)$ be the structuring element, which is the discrete function over a domain $B(m)=\{0,1.....,M-1\}$ two basic morphological operators, the erosion and the dilation, can be defined as

$$(f\Theta B)(n)=\underset{m=0,...M-1}{MIN}\{f(n+m)-B(m)\} \quad (1)$$

$$(f\oplus B)(n)=\underset{m=0,...M-1}{MAX}\{f(n-m)+B(m)\} \quad (2)$$

Based on de dilation and erosion, two other basic morphological operators, the opening ($\circ$) and the closing ($\bullet$) can be further defined:

$$(f\circ B)(n)=(f\Theta B\oplus B)(n) \quad (3)$$

$$(f\bullet B)(n)=(f\oplus B\Theta B)(n) \quad (4)$$

Table 1 and 2 illustrate successively the basic properties of dilation and erosion, closing and opening operations.

To obtain the eroded function of $f(n)$, we attribute to $f(n)$ its minimal value in the field of the structuring element $B(m)=(0,0,0,0,0)$ which is a line segment, and with each new displacement of $B(m)$, the structuring element $B(m)$ plays the same role as a moving window. The width of such window is chosen empirically $B=5$. This is illustrated in Fig.1, where these operations are applied to an ECG signal.

The illustration shows a reduction in the peaks of ECG signal and a widening of the valleys. Erosion is an operator of shrinking in which the values of $f\Theta B$ are always less than those of f.

Dilataion
Commutative: $A\oplus B=B\oplus A$
Associative: $(A\oplus B)\oplus C=A\oplus(B\oplus C)$
Increasing: $(A)_x\oplus B=(A\oplus B)_x$
Duality: $A\oplus B=(A^c\ominus B)^c$

Erosion
Non-commutative : $A\ominus B\neq B\ominus A$
Translation Invariance : $(A)_x\ominus B=(A\ominus B)_x$
Increasing: $B_1\subseteq B_2\Rightarrow A_1\ominus B_2\subseteq A\oplus B_1$
Duality: $(A\ominus B^c)^c=A^c\oplus B$

Table 1 : Basic properties of dilation and erosion

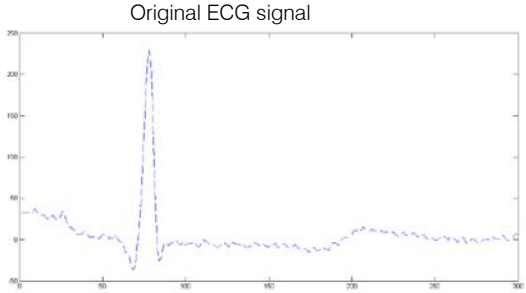
Closing
Extensivity: $A\subseteq A\bullet B$
Idempotence: $(A\bullet B)\bullet B=A\bullet B$
Translation Invariance: $:(A)_x\bullet B=(A\bullet B)_x$
Increasing: $A_1\subseteq A_2\Rightarrow A_1\bullet B\subseteq A_2\bullet B$
Duality: $(A\bullet B)^c=AoB$

Opening
Antiextensivity $AoB\subseteq A$
Idempotence: $(AoB)oB=AoB$
Translation Invariance: $:(A)_xoB=(AoB)_x$
Increasing: $A_1\subseteq A_2\Rightarrow A_1oB\subseteq A_2oB$
Duality: $(AoB)^c=A\bullet B$

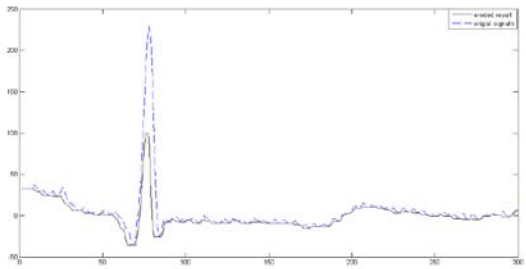
Table 2 : Basic properties of closing and opening

Similarly, the dilation can be performed by taking of set sums. Its complexity is the same as erosion and is related to convolution, where instead of doing summation of products, a maximum of sums is computed.

Original ECG signal



Erosion of *B* by *f*



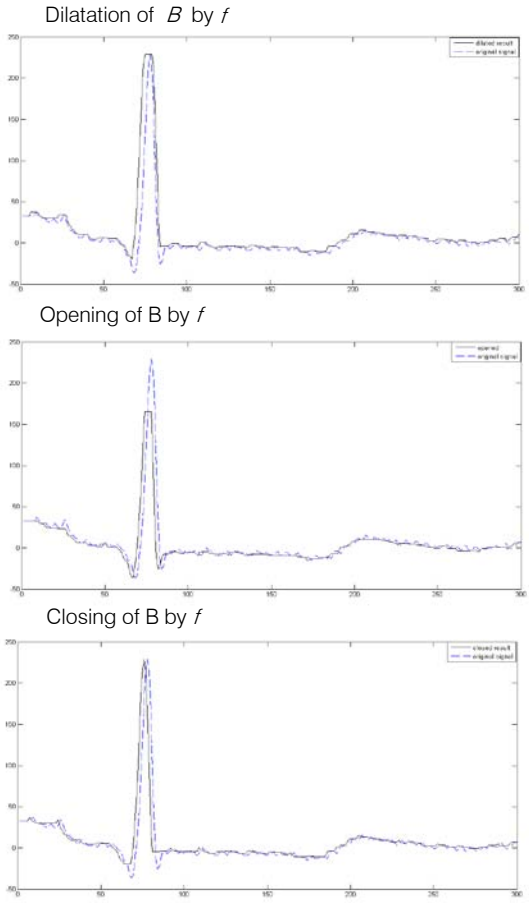


Fig. 1 : Erosion, Dilation, Opening, Closing of ECG signal by structuring element

This transformation fills the valleys and thickens the peaks. Fig. 1 shows that dilation is an operation of expansion in which values of $f \oplus B$ are always greater than those of f .

Morphological opening can be expressed as an erosion operation followed by a dilation of the eroded result, using the same structuring element. The dual operation of opening is the closing operation.

Morphological closing can be expressed as a dilation operation followed by an erosion of the dilated result, using the same structuring element.

Figure1. shows that the opening by B smooths the graph off from below by cutting down its peaks.

The closing smooths the graph of f from above by filling up its valleys (suppress peaks). Subtracting from f its opening or closing by B provides respectively the peaks and valleys of f .

The width of these peaks and valleys depends on the size of B . Therefore, opening and closing by a structuring element B can be used effectively to suppress noise in ECG signals.

b) Morphological structuring element (SE)

After selecting the morphological operator, the SE is the next key component of the morphology analysis to be defined. Generally, only when the shape of the signal is matched to those of SE, the signal can be preserved. Therefore, the shape, length (domain) and height (amplitude) of SE should be selected according to the signal to be analyzed.

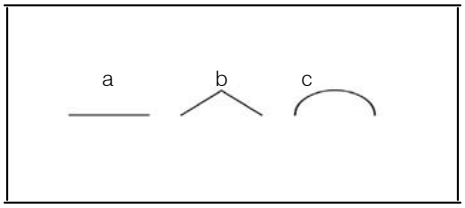


Fig. 2 : Structuring element E of various shapes: (a) flat SE; (b) triangular SE; and (c) semicircular SE

The shapes of SE can vary from regular to irregular curves, such as flat, triangle, semicircle, and so on figure 2. illustrates some of the common SE.

III. Algorithm of the Morphological Top-Hat Transform for ECG Signal

In the morphological Top-Hat transform algorithm, the noise suppression is performed as follows (20):

$$\begin{aligned} f &= f_o \bullet B - f_o \circ B \\ &= (f_o \bullet B - f_o) + (f_o - f_o \circ B) \\ &= (f_o \oplus B_1 \Theta B_2 - f_o) + (f_o - f_o \Theta B_1 \oplus B_2) \end{aligned} \quad (5)$$

$f_o \bullet B - f_o$ and $f_o - f_o \circ B$ are two types of the morphological Top-Hat Transform [21]. The morphological Top-Hat transform is a high-pass filter with good performances. $f_o \bullet B - f_o$ is called the Black Top-Hat transform, which is used to extract negative impulsive features; $f_o \circ B - f_o$ is called the White Top-Hat transform, which is used to extract positive impulsive features. Thus filter can be used to extract the positive and negative features simultaneously. Figure3. illustrates a bloc diagram describing the structure of the morphological Top-Hat transform of the ECG signals. It consists of three blocs: The first is concerned with the acquisition of ECG signals (f_o : original ECG signal). This step is followed by another step which allows the detection of the noise. This detection is achieved using the morphological operators defined in equation (5). B_1 and B_2 are structuring elements for opening and closing. These operations are used simultaneously on the original signal. The following step is the subtraction of the resulted closing and opening operations. This filtered

ECG signal f is the resultant signal after noise suppression.

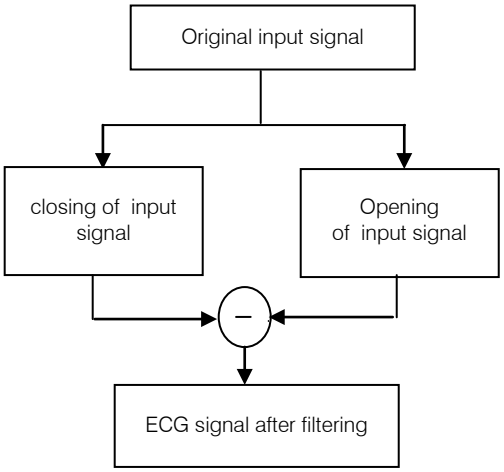


Fig. 3 : Bloc diagram of Top-Hat transform

The $B_{pair}(B_1, B_2)$ is selected according to the purpose of analysis and to the morphological properties of the ECG signal. B_1 is selected to be a triangular shape to retain the peaks and valleys and B_2 is a line segment to remove the noise.

IV. Results and Discussion

We use the MIT-BIH arrhythmia to evaluate the morphological Top-Hat algorithm. All the programs are written in Matlab 7 environment under the platform Pentium 4.

After the acquisition of ECG signal, the following stage is the suppression of the noise. It consists on the application of operators of Morphological Top-Hat transform. In fact, the input signal simultaneously is processed by the operations "closing" and "opening", followed by a subtraction, to generate at the end the filtered signal. Thus, the Top-Hat transform can be used to extract the positive and negative features simultaneously.

It should be noted that the shape of the structuring element in the suppression of noise is different compared to that from the correction of the baseline. Indeed, it can take two different forms of equal lengths: a triangular form B_1 to maintain the peaks and the valleys on a straight form (segment of null amplitude) B_2

In our case the size of the structuring element was fixed at 5 samples units each, with values of $B_1 = (0,1,5,1,0)$, $B_2 = (0,0,0,0,0)$. This value is fixed in an empirical way where the minimum and the maximum are fixed at optimal values in the stage of the suppression of the noise. The algorithm is applied respectively to the records 101, 105, 108 and 209 of the MIT-BIH data base. As shown in Figures. (4-5-6-7), good performance of suppression of the noise can be observed.

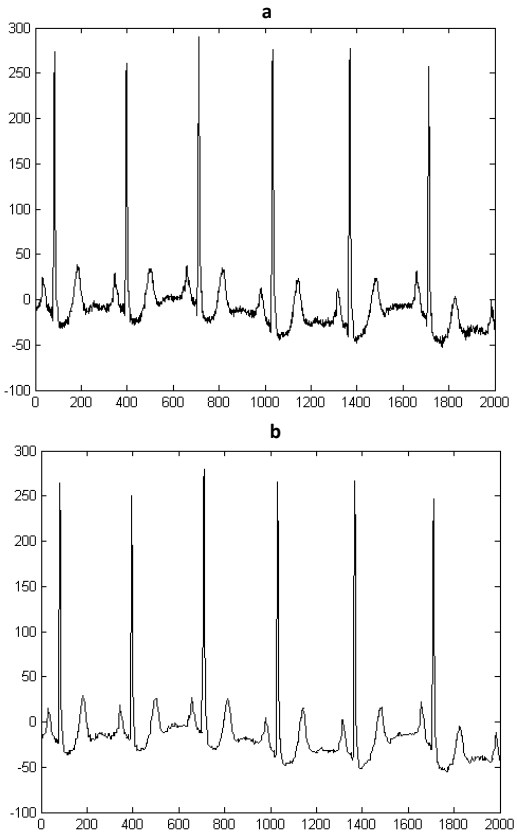
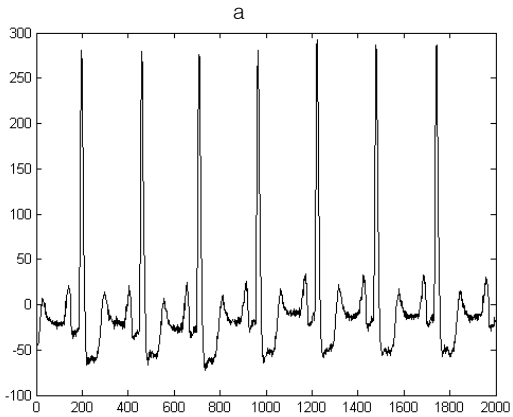


Fig. 4 : Results of MTHT (a) original ECG signal 101; (b) denoised signal



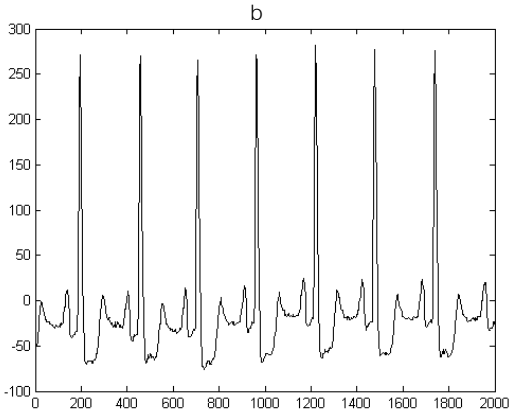


Fig. 5 : Results of MTHT (a) original ECG signal 105; (b) denoised signal

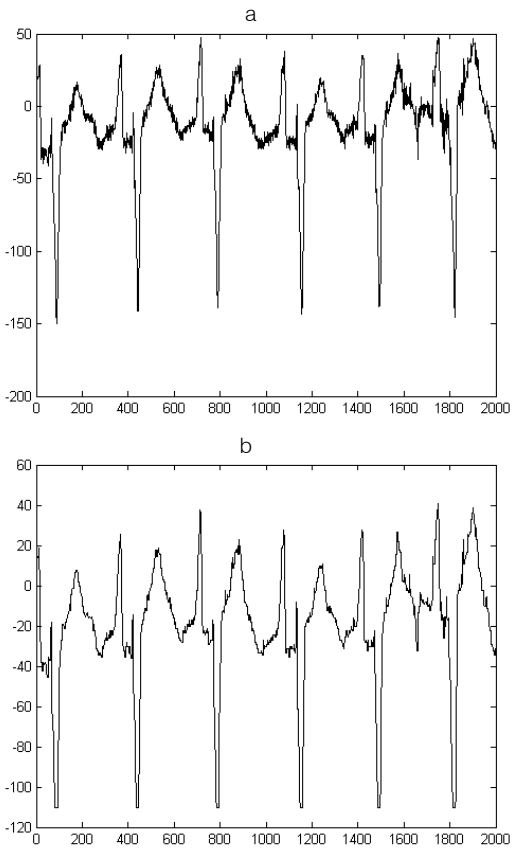


Fig. 6 : Results of MTHT (a) original ECG signal 108; (b) denoised signal

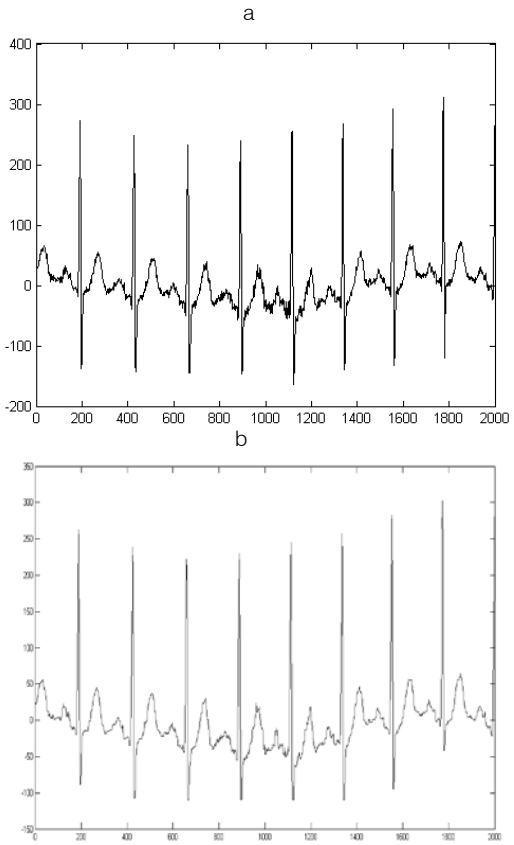


Fig. 7 : Results of MTHT (a) original ECG signal 209; (b) denoised signal

a) Simulation Study

In order to test the performance of the developed Top-Hat transform algorithm in suppressing noise, the algorithm is applied to ECG signal recordings of the data base "MIT-BIH Arrhythmia Database" to which simulated Gaussian noise of a 5 dB level has been added.

The resulting test signal is given by:

$$S(n) + f(n) = f_S(n) \qquad (6)$$

Where $S(n)$ is the simulated Gaussian noise, $f(n)$ is the ECG signal recording and $f_S(n)$ is the noisy ECG signal.

Figures 8. upto 14 illustrats the results obtained after applying the proposed algorithm a noisy ECG signal recordings. As it can be noticed in each figure, the denoised ECG signals (c) are well resolved mainly the different waves, and the added noise is completed suppressed.

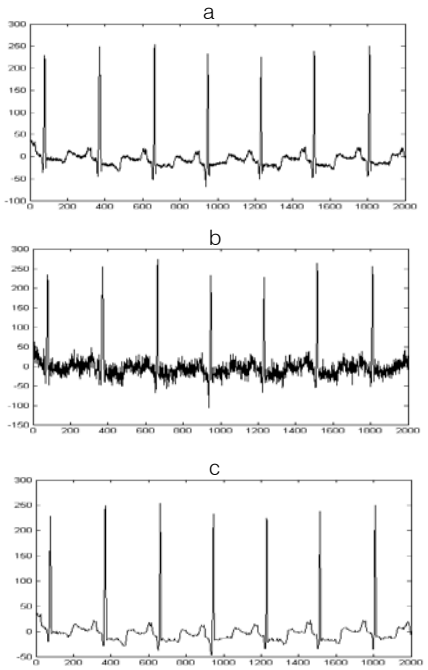


Fig. 8 : Application of the proposed denoising algorithm to a 5 dB noisy normal sinus rhythm ECG signal record 100. a) Input ECG signal; b) with added Gaussian noise ECG signal; c) denoised ECG signals.

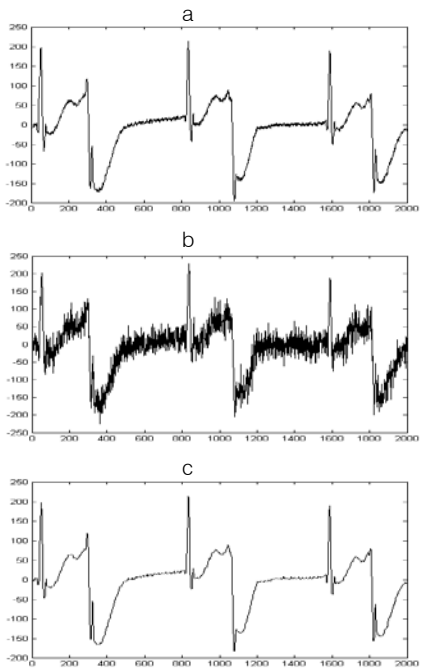


Fig. 10 : Application of our denoi7 a) Input ECG signal; b) with added Gaussian noise EC sing algorithm to a 5 dB noisy ECG signal with a PVC record 20G signal; c) denoised ECG signals.

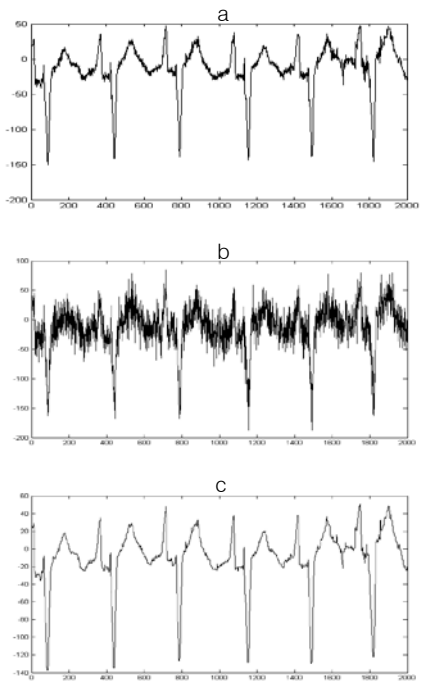


Fig. 9 : Application of our denoising algorithm to a 5 dB noisy ECG signal with a PVC record 108 a) Input ECG signal; b) with added Gaussian noise ECG signal; c) denoised ECG signals.

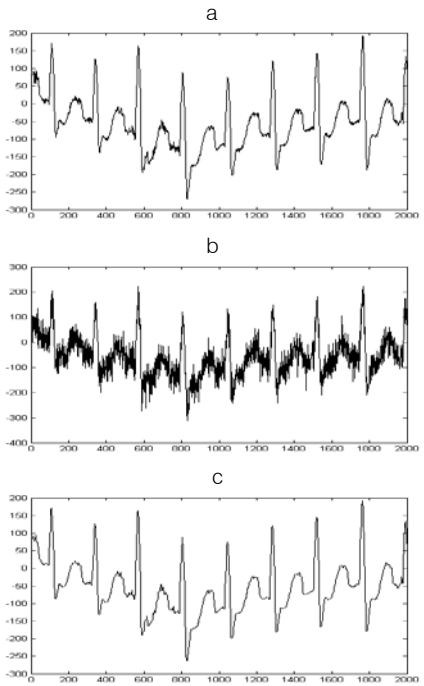


Fig. 11 : Application of our denoising algorithm to a 5 dB noisy ECG signal containing a baseline wandering record 109: a) Input ECG signal; b) with added Gaussian noise ECG signal; c) denoised ECG signals.

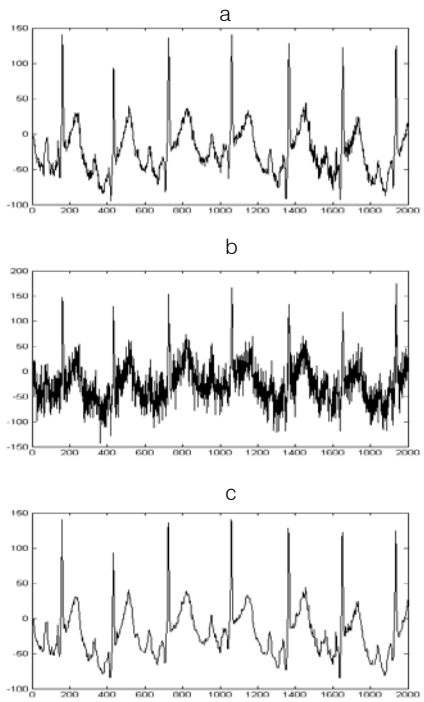


Fig. 12 : Application of our denoising algorithm to a 5 dB noisy ECG signal containing a baseline wandering record 228: a) Input ECG signal; b) with added Gaussian noise ECG signal; c) denoised ECG signals.

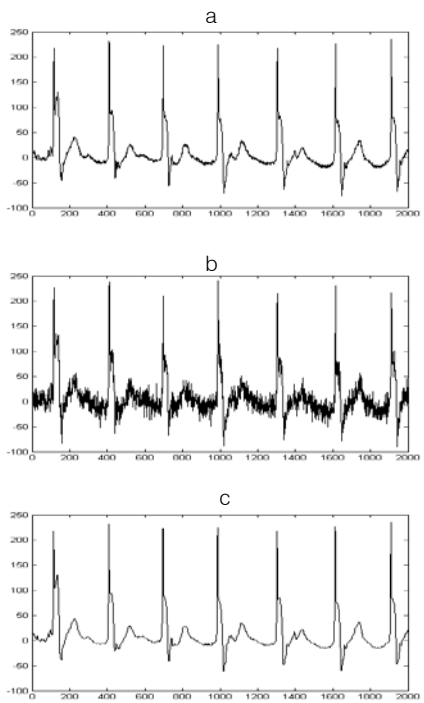


Fig. 13 : Application of our denoising algorithm to a 5 dB noisy ECG signal containing a various waves record.

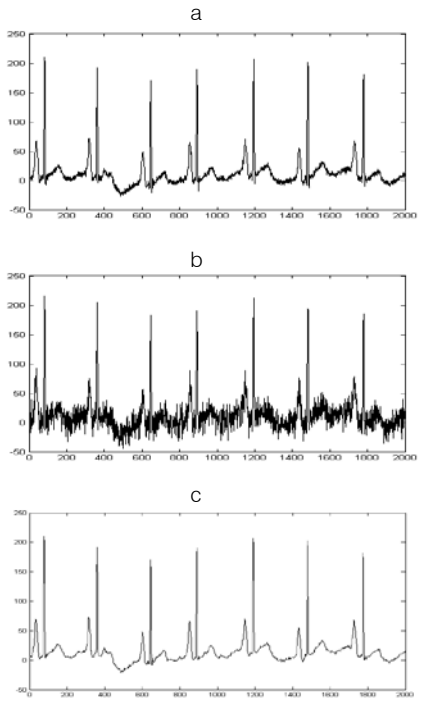


Fig. 14 : Application Application of our denoising algorithm to a 5 dB noisy ECG signal containing a various waves record 222: a) Input ECG signal; b) with added Gaussian noise ECG signal; c) denoised ECG signals.

b) Evaluation criteria values

General comparative study was related to three recordings of the database ` MIT-BIH Arrhythmia Database' which are ` 100.dat ', ` 101.dat ', and ` 103.dat '.

Tables 3 and 4 summarize the values of signal to noise ratio SNR (eq.7) and the Mean Squared error MSE (eq.8), respectively, of the application of three approaches of filtering (Top-Hat transform, the wavelet denoising utilize the Visu Shrink, and low pass Butterworth filter [22-23] to each recording.

Table 3: The obtained output SNR (in dBs) values for each tested filtering method applied to the set of MIT-BIH arrhythmia Databases ECG records ('100.dat', '101.dat' and '103.dat').

Table 4: The obtained MSE (in dBs) values for each tested filtering method applied to the set of MIT-BIH arrhythmia Databases ECG records ('100.dat', '101.dat' and '103.dat').

One note that the average values of the average quadratic error and the signal report/ratio on noise are, in fact, calculated for various successive segments of 4096 samples.

$$SNR = 10\log \frac{\sum_{n=1}^{N}(f_s(n))^2}{\sum_{n=1}^{N}(f_s(n) - f_d)^2} \quad (7)$$

$$MSE = \frac{\sum_{n=1}^{N}(f_s(n) - f_d(n))^2}{N} \times 100 \quad (8)$$

Where f_s denotes the reference ECG signal and f_d represents the constructed denoised ECG signal. Whereas N is the length of the data segment.

Values SNR and MSE obtained of the general comparative study (of tables 3 and 4) show, the superiority of performance of the algorithm Top-Hat transform.

Table 5: Presents the SNR between the denoised ECG signal and the clean ECG signal, as obtained for each method. The ECG signal was corrupted by adding white Gaussian noise. The noise level was adjusted to specific values in such a way as to obtain the different SNRs. Top-Hat transform was used by the proposed algorithm in Table 5.

Table 6: Lists the SNR of the proposed method and the other methods for the ECG signals (MIT-BIH database record 103) corrupted with electrode motion artifact noise.

Table 7: Presents the SNR of the proposed method and the other methods for the ECG signals (MIT-BIH database record 103) corrupted with muscle artifact noise.

Table 8: Lists the SNR of the proposed method and the others for the ECG signals (MIT-BIH database record 109) corrupted with electrode motion noise.

The results clearly indicate that the proposed method has stronger denoising abilities than the others methods. Although the other methods removed the

Table 3 : The obtained output SNR (in dBs) values for each tested filtering method applied to the set of MIT-BIH arrhythmia Databases ECG records ('100.dat', '101.dat' and '103.dat').

Method record	Input SNR (dB)	Top-Hat transform	Low pass Butterworth filer	VisuShrink
100 dat	5	$0.0005*10^{-5}$	$0.0015*10^{-5}$	$0.00098*10^{-5}$
101 dat	5	$0.0002*10^{-4}$	$0.0016*10^{-4}$	$0.0011*10^{-4}$
103 dat	5	$0.008*10^{-4}$	$0.0016*10^{-4}$	$0.0012*10^{-4}$

102 : a) Input ECG signal; b) with added Gaussian noise ECG signal; c) denoised ECG signals

Table 4 : The obtained MSE (in dBs) values for each tested filtering method applied to the set of MIT-BIH arrhythmia Databases ECG records ('100.dat', '101.dat' and '103.dat').

Method record	Input SNR (dB)	Top-Hat transform	Low pass Butterworth filer	VisuShrink
100 dat	5	13.2025	5.4718	4.8539
101 dat	5	12.5645	6.7585	6.8030
103 dat	5	12.9845	9.1075	7.9001

Table 5 : The performance (SNR) of the proposed method and others for the ECG signal (MIT-BIH database record 103) corrupted with Gaussian noise.

Input SNR	Output SNR Proposed method	Output SNR for VisuShrink	Output SNR for SureShrink	Output SNR for ayesShrink
6.8	12.08	10.50	10.82	10.54
9.29	13.88	12.73	13.24	12.88
12.81	17.80	15.84	16.57	15.99
15.83	20.73	18.31	19.25	18.33

Table 6 : The performance (SNR) of the proposed method and others for the ECG signal (MIT-BIH database Record 103) corrupted with electrode motion noise.

Input SNR	Output SNR Proposed method	Output SNR for VisuShrink	Output SNR for SureShrink	Output SNR for ayesShrink
6.8	11.30	9.45	10.85	10.72
9.29	13.45	11.65	13.27	13.05
12.81	16.70	15.74	16.60	16.14
15.83	19.60	17.34	19.31	18.48

Table 7 : The performance (SNR) of the proposed method and others for the ECG signal (MIT-BIH database record 103) corrupted with muscle artifact noise.

Input SNR	Output SNR Proposed method	Output SNR for VisuShrink	Output SNR for SureShrink	Output SNR for ayesShrink
6.8	13.20	11.15	12.88	12.68
9.29	15.76	13.24	15.26	14.91
12.81	18.92	16.28	18.48	17.79
15.83	21.49	18.74	21.02	19.85

Table 8 : The performance (SNR) of the proposed method and others for the ECG signal (MIT-BIH database record 109) corrupted with electrode motion noise.

Input SNR	Output SNR Proposed method	Output SNR for VisuShrink	Output SNR for SureShrink	Output SNR for ayesShrink
6.8	10.96	10.57	10.91	10.85
9.29	13.33	12.77	13.26	13.22
12.81	16.77	15.76	16.65	16.62
15.83	19.80	18.25	19.28	19.19

Noise, the signal was distorted, as well. Our method showed good performance with a Top-Hat transform. The values obtained from our proposed method were always among the ones that had the highest SNR.

V. Conclusion

A new algorithm Morphological Top-Hat transform for noise suppression using nonlinear transform was developed and evaluated. It consists morphological operators which are closing and opening. These operators are used as structuring element SE. Such SE was chosen as a triangular shape to maintain the peaks and the valleys on a straight form. The algorithm was implemented and evaluated a same ECG signals for the MIT-BIH data base. The experimental application of this algorithm result is better than Visu Shrink, Sure Shrink, Bayes Shrink and Low passes Butterworth filer, particularly, in ECG signal denoising.

References Références Referencias

1. P. Laguna, R. Jané, and P. Caminal, "Automatic detection of wave boundaries in multilead ECS signals: validation with the CSE Database", Computres and Biomedical 1994, Vol. 27, pp 45-60.
2. D. Vel, O.Y, R-wave detection in the presence of muscle artifacts, IEEE Trans. Bio.Eng, 1984, vol. 31, pp 751-717.
3. S. A. TAOULI, F. BEREKSI REGUIG, "Detection of QRS Complexes in ECG Signals Based on Empirical Mode Decomposition", Global Journal of Computer Science and Technology Volume 11 Issue 20 Version December 2011.
4. M. Boutaa, "Analyze and quantification of the correlation of the rate of heartbeat with the various components of ECG signal", Thesis of Magister in Electronics, University Aboubekr Belkaid, Biomedical Electronics Department, June 2006.
5. L. Zhang, A. bao, "Edge detection by scale multiplication in wavelet domain", Pattern Recognition Letters, 2002, Vol. 23, no. 14, pp 1771-1784.
6. L. Zhang, A. bao, and X. Wu, "Canny edge detection enhacement by scale multiplication", IEEE Trans. On Pattern Analysis and Machine Intelligence, 2005, Vol. 27, no. 9, pp 1485- 1490.
7. D. L. Donoho, "De-noising by soft-thresholding", IEE Trans. Inform. Theory 1992; 41(3), pp 612-627.
8. Y. Xu, J.B. Weaver, D.M. Healy, and J. Lu, "Wavelet transform domain filters: a spatially selective noise filtration technique", IEEE Trans.
9. Image Processing, 1994, Vol. 3, no. 6, pp 747-758.D.L. Donoho, I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage", J. Amer. Statist. Assoc., 1995, Vol. 90, pp 1200-1224.
10. S. Chang, B. Yu, M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression", IEEE Trans. Image Processing, 2000, Vol. 9, pp 1532-1546.
11. D.L. Donoho, I.M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage", Biometrika, 1994, Vol. 81, pp 425-455.
12. C.M. Stein, "Estimation of the mean of amultivariate normal distribution", Ann. Statist, 1981, Vol. 9, pp 1135-1151.
13. B.N. Singh, A.K. Tiwari, "Optimal selection of wavelet basis function applied to ECG signal denoising", Digital Signal Processing, 2006, Vol. 16, pp 275-287.
14. J. Serra, Cours of mathematical morphology, Mathematical Morphology Center, School of the Mines in Paris, 2000.
15. http://malte.ensmp.fr/~serra//cours/.MIT-BIH, Arrhythmia Database Directory, Harvard MIT Division of Health Sciences and Technology, Biomes. Eng. Center, 1997.
16. S. A. TAOULI, F. BEREKSI REGUIG, "Application of the morphological filter in the processing of the electrocardiogram signal ECG", CNIEE 04, 22-23, Oran, Algeria, 2004.
17. P. Maragos, R.W. Schafer, "Morphological filters part I and II", IEEE Trans. Acoust. Speech Signal Process. ASSP-1987, Vol. 35, pp 1170-1184.
18. Y. Sun, K.L. Chan, S.M. Krishnan, "Characteristic wave detection in ECG using morphological transform", BioMed Central Cardiovasculair 2005, 5:28, pp 1-7.
19. S.A TAOULI, F. BEREKSI REGUIG, "Application of the operator's morphology for the detection of the waves the ECG signal", International conference on Electrical engineering ICEE, Oran, Algeria, 2006.
20. S. A. Taouli, "Analyzes variability of interval QT and its correlation with the cardiac rhythm of the electrocardiogram signal ECG", Doctorate thesis, Biomedical Electronics Department, Abou Bekr

Belkaid University-Tlemcen, Tlemcen, Algeria January 2013.
21. F. Meyer, “Iterative image transformations for an automatic screening of cervical smears”, Journal of Histochemistry and Cytochemistry, 1979, 27 (1), pp 128–135.
22. S. A. Chouakri, “treatment and analysing ECG signal by wavelets transform”, Thesis of Doctorate in Electronics (Option: Signals an Systems), University Aboubekr Belkaid, Biomedical Electronics Department, December 2007, pp 71-102.
23. S. A. Chouakri, F. Bereksi-Reguig, “ECG smoothing based on combining wavele denoising levels”, IEE Asian Journal of Information Technology, 2006, 5(6), pp 666-677.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 5 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Audio Compression using Munich and Cambridge Filters for Audio Coding with Morlet Wavelet

By S. China Venkateswarlu, V. Sridhar, A. Subba Rami Reddy & K. Satya Prasad

Holy Mary Institute of Technology & Science

Abstract - The main aim of work is to develop morlet wavelet using Munich and Cambridge filters, for audio compression and most psycho-acoustic models for coding applications use a uniform -equal bandwidth, spectral decomposition for compression. In this paper we present a new design of a psycho-acoustic model for audio coding following the model used in the standard MPEG-1 audio layer 3. This architecture is based on appropriate wavelet packet decomposition instead of a short term Fourier transformation. To fulfill this aim, the following objectives are carried out: Approximate the frequency selectivity of the human auditory system. However, the equal filter properties of the uniform sub- bands do not match the non uniform characteristics of cochlear filters and reduce the precision of psycho-acoustic modeling. This architecture is based on appropriate wavelet packet decomposition instead of a short term Fourier transformation. In this paper Morlet Munich coder shows best performance. The MPEG/Audio is a standard for both transmitting and recording compressed ratio. The MPEG algorithm achieves compression by exploiting the perceptual limitation of the human ear.

Keywords : *audio compression, acoustic, human voice, morlet wavelet, munich, psycho-acoustic, music, wavelet transform.*

GJCST-C Classification : H.5.5

Strictly as per the compliance and regulations of:

Audio Compression using Munich and Cambridge Filters for Audio Coding with Morlet Wavelet

S. China Venkateswarlu [α], V. Sridhar [σ], A. Subba Rami Reddy [ρ] & K. Satya Prasad [ω]

Abstract - The main aim of work is to develop morlet wavelet using Munich and Cambridge filters, for audio compression and most psycho-acoustic models for coding applications use a uniform -equal bandwidth, spectral decomposition for compression. In this paper we present a new design of a psycho-acoustic model for audio coding following the model used in the standard MPEG-1 audio layer 3. This architecture is based on appropriate wavelet packet decomposition instead of a short term Fourier transformation. To fulfill this aim, the following objectives are carried out: Approximate the frequency selectivity of the human auditory system. However, the equal filter properties of the uniform sub- bands do not match the non uniform characteristics of cochlear filters and reduce the precision of psycho-acoustic modeling. This architecture is based on appropriate wavelet packet decomposition instead of a short term Fourier transformation. In this paper Morlet Munich coder shows best performance. The MPEG/Audio is a standard for both transmitting and recording compressed ratio. The MPEG algorithm achieves compression by exploiting the perceptual limitation of the human ear. Audio compression algorithms are used to obtain compact digital representations of high-fidelity audio signals for the purpose of efficient transmission. The main objective in audio coding is to represent the signal with a minimum number of bits while achieving transparent signal reproduction. The majority of MPEG1coders apply a psycho-acoustic model for coding applications using filter bank to approximate the frequency selectivity of the human auditory system. The ear is an organ of large sensibility, presenting a high resolution and a great dynamic of the signal. The ear is more sensitive to low frequencies than to high ones and our hearing threshold is very intense in the high frequency regions. The study of its characteristics makes it possible to exploit the intrinsic properties of the ear in order to carry out systems of compression with loss of information. Most MP3 coders use the Fourier transform whose fundamental roles are limited only to the spectral properties of the signal. In this paper, a new design for modeling auditory masking is based on wavelet packet decomposition.

Keywords : *audio compression, acoustic, human voice, morlet wavelet, munich, psycho-acoustic, music, wavelet transform.*

Author α : Research Scholar JNTUK-Kakinada.
E-mail : cvenkateswarlus@gmail.com
Author σ : Asst. Professor ECE, VJIT, Principal SKIT, Tirupati.
Author ρ : Rector JNTUK-Kakinada.

I. Introduction

Audio compression is a form of data compression designed to reduce the transmission bandwidth requirement of digital audio_streams and the storage size of audio files. Audio compression algorithms are implemented in computer software as audio codecs. Generic data compression algorithms perform poorly with audio data, seldom reducing data size much below 87% from the original and are not designed for use in real time applications. Consequently, specifically optimized audio lossless and lossy algorithms have been created. Lossy algorithms provide greater compression rates and are used in mainstream consumer audio devices.In both lossy and lossless compression, information__redundancy is reduced, using methods such as coding, pattern recognition and linear prediction to reduce the amount of information used to represent the uncompressed data. The trade-off between slightly reduced audio quality and transmission or storage size is outweighed by the latter for most practical audio applications in which users may not perceive the loss in playback rendition quality. For example, one Compact Disc holds approximately one hour of uncompressed high fidelity music, less than 2 hours of music compressed losslessly, or 7 hours of music compressed in the MP3 format at medium bit rates.

a) Lossless Audio Compression

Lossless audio compression produces a representation of digital data that can be expanded to an exact digital duplicate of the original audio stream. This is in contrast to the irreversible changes upon playback from lossy compression techniques such as Vorbis and MP3. Compression ratios are similar to those for generic lossless data compression (around 50–60% of original size, and substantially less than for lossy compression, which typically yield 5–20% of original size.

b) Lossless Compressed Audio Formats

A lossless compressed format requires much more processing time than an uncompressed format but is more efficient in space usage. Uncompressed audio formats encode both sound and silence with the same

number of bits per unit of time. Encoding an uncompressed minute of absolute silence produces a file of the same size as encoding an uncompressed minute of symphonic orchestra music. In a lossless compressed format, however, the music would occupy a marginally smaller file and the silence take up almost no space at all. Lossless compression formats (such as the most widespread FLAC, WavPack, Monkey's Audio, ALAC/Apple Lossless) provide a compression ratio of about 2:1. Development in lossless compression formats aims to reduce processing time while maintaining a good compression ratio.

II. Audio Signal Processing

Audio signal processing, sometimes referred to as audio processing, is the intentional alteration of auditory signals, or sound. As audio signals may be electronically represented in either digital or analog format, signal processing may occur in either domain. Analog processors operate directly on the electrical signal, while digital processors operate mathematically on the digital representation of that signal. Audio processing was necessary for early radio broadcasting-as there were many problems with studio to transmitter links. *Analog signals*-An analog representation is usually a continuous, non-discrete, electrical; a voltage level represents the air pressure waveform of the sound. *Digital signals*- A digital representation expresses the pressure wave-form as a sequence of symbols, usually binary numbers. This permits signal processing using digital circuits such as microprocessors and computers. Although such a conversion can be prone to loss, most modern audio systems use this approach as the techniques of digital signal processing are much more powerful and efficient than analog domain signal processing.

a) Sound Recording and Reproduction

Sound recording and reproduction is an electrical or mechanical inscription and re-creation of sound waves, such as spoken voice, singing, instrumental music, or sound effects. The two main classes of sound recording technology are analog recording and digital recording. Acoustic analog recording is achieved by a small microphone diaphragm that can detect changes in atmospheric pressure (acoustic sound waves) and record them as a graphic representation of the sound waves on a medium such as a phonograph (in which a stylus senses grooves on a record). In magnetic tape recording, the sound waves vibrate the microphone diaphragm and are converted into a varying electric current, which is then converted to a varying magnetic field by an electromagnet, which makes a representation of the sound as magnetized areas on a plastic tape with a magnetic coating on it. Analog sound reproduction is the reverse process, with a bigger loudspeaker diaphragm causing changes to atmospheric pressure to form acoustic sound waves. Electronically generated sound waves may also be recorded directly from devices such as an electric guitar pickup or a synthesizer, without the use of acoustics in the recording process other than the need for musicians to hear how well they are playing during recording sessions.

Digital recording and reproduction converts the analog sound signal picked up by the microphone to a digital form by a process of digitization, allowing it to be stored and transmitted by a wider variety of media. Digital recording stores audio as a series of binary numbers representing samples of the amplitude of the audio signal at equal time intervals, at a sample rate so fast that the human ear perceives the result as continuous sound. Digital recordings are considered higher quality than analog recordings not necessarily because they have higher fidelity (wider frequency response or dynamic range), but because the digital format can prevent much loss of quality found in analog recording due to noise and electromagnetic interference in playback, and mechanical deterioration or damage to the storage medium. A digital audio signal must be reconverted to analog form during playback before it is applied to a loudspeaker or earphones.

b) Speech Compression

Speech compression may mean different things: Speech encoding refers to compression for transmission or storage, possibly to an unintelligible state, with decompression used prior to playback. Time-compressed speech refers to voice compression for immediate playback, without any decompression (so that the final speech sounds faster to the listener). Wavelet unlike FFT, whose basic functions are sinusoids, is based on small waves, called wavelets, of varying frequency and limited duration. Its most important characteristics are conceived to analyze temporal and spectral properties of non-stationary signals such as audio. A new architecture of a psycho-acoustic model based on wavelets decomposition can be applied to the audio coders in sub-bands approximating the critical bands. Models of human auditory pathway are commonly used as a part of audio coding systems or algorithms for objective Evaluation of sound quality. In order to properly simulate the nonlinear, signal-dependent behavior of the cochlea, physiologically accurate models were developed, such as the Munich model which was concepted by Zwicker and as for the Cambridge one by Moore. Our study is made up of fourth parts. The first part will focus on the proposed psycho-acoustic model. The second part aims at presenting the two models (Munich and Cambridge). The third part deals with a comparison between the compressions ratios of the proposed coders (FFT and wavelet). Finally, a sound quality evaluation will be carried out to conclude this work. The

MPEG/Audio is a standard for both transmitting and recording compressed ratio. The MPEG algorithm achieves compression by exploiting the perceptual limitation of the human ear. Audio compression algorithms are used to obtain compact digital representations of high-fidelity audio signals for the purpose of efficient transmission. The main objective in audio coding is to represent the signal with a minimum number of bits while achieving transparent signal reproduction. The majority of MPEG1coders apply a psycho-acoustic model for coding applications using filter bank to approximate the frequency selectivity of the human auditory system. The ear is an organ of large sensibility, presenting a high resolution and a great dynamic of the signal. The ear is more sensitive to low frequencies than to high ones and our hearing threshold is very intense in the high frequency regions. The study of its characteristics makes it possible to exploit the intrinsic properties of the ear in order to carry out systems of compression with loss of information. Most MP3 coders use the Fourier transform whose fundamental roles are limited only to the spectral properties of the signal. In this paper, a new design for modelling auditory masking is based on wavelet packet decomposition. Wavelet unlike FFT, whose basic functions are sinusoids, is based on small waves, called wavelets, of varying frequency and limited duration. Its most important characteristics are conceived to analyze temporal and spectral properties of non-stationary signals such as audio. A new architecture of a psycho-acoustic model based on wavelets decomposition can be applied to the audio coders in sub-bands approximating the critical bands. Models of human auditory pathway are commonly used as a part of audio coding systems or algorithms for objective Evaluation of sound quality. In order to properly simulate the nonlinear, signal-dependent behaviour of the cochlea, physiologically accurate models were developed, such as the Munich model which was concepted by Zwicker and as for the Cambridge one by Moore. Our study is made up of fourth parts. The first part will focus on the proposed psycho-acoustic model. The second part aims at presenting the two models (Munich and Cambridge). The third part deals with a comparison between the compression ratios of the proposed coders (FFT and wavelet). Finally, a sound quality evaluation will be carried out to conclude this work.

III. Wavelet Transform

Given a real, mother wavelet function $\psi(t) \in L^2(R)$, which satisfies the admissibility condition,

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty \quad (3.1)$$

Where $\Psi(\omega)$ is the Fourier transform of $\psi(t)$, the scaling parameter p and the shifting parameter q are introduced to construct a family of basis functions or wavelets $\{\psi_{p,q}(t)\}, p,q \in R, p \neq 0$, where

$$\psi_{p,q}(t) = \frac{1}{\sqrt{|p|}} \psi\left(\frac{t-q}{p}\right) \quad (3.2)$$

If $\Psi(\omega)$ has sufficient decay, the admissibility condition reduces to the constraint that $\Psi(0) = 0$ or

$$\int_{-\infty}^{\infty} \psi(t)dt = \Psi(0) = 0 \quad (3.3)$$

Because the Fourier transform is zero at the origin, and the spectrum decays at high frequencies, the wavelet behaves as a bandpass filter. The continuous wavelet transform (CWT) of a function or a signal $x(t) \in L^2(R)$ is then defined as

$$X_{CWT}(p,q) = \int_{-\infty}^{\infty} x(t)\psi_{p,q}(t)dt \quad (3.4)$$

and the inverse wavelet transform is given by

$$x(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X_{CWT}(p,q)\psi_{p,q}(t)\frac{dpdq}{p^2} \quad (3.5)$$

Since $\{\psi_{p,q}(t)\}$ is a redundant basis set, it implies that, by discretizing the scaling and shifting parameters as follows:

$$p = 2^k, q = 2^k l,\ k,l \in Z \quad (3.6)$$

an orthonormal basis $\{\psi_{k,l}(t)\}$ is obtained,

$$\psi_{k,l}(x) = 2^{-k/2}\psi(2^{-k}x - l),\ k,l \in Z \quad (3.7)$$

which comply with the condition,

$$\int_{-\infty}^{\infty} \psi_{k,l}(t)\psi_{k',l'}(t)dt = \delta(k-k')\delta(l-l'), \forall k,k',l,l' \in Z \quad (3.8)$$

The wavelet transform is an appropriate tool for multiresolution analysis (MRA) [47]. In MRA, a *scaling function* $\phi(x)$ and the associated mother wavelet function $\psi(x)$, which are needed in the

construction of a *complete basis* (see below), must satisfy the two-scale difference equations:

$$\phi(x) = \sqrt{2}\sum_n h(n)\phi(2x-n)$$
$$\psi(x) = \sqrt{2}\sum_n g(n)\phi(2x-n) \quad (3.9)$$

Where the coefficients $h(n)$ and $g(n)$ satisfy the following:

$$g(n) = (-1)^n h(1-n) \quad (3.10)$$

Similar to the construction of $\{\psi_{k,l}(t)\}$, a family of orthonormal basis $\{\phi_{k,l}(x)\}$ can be obtained through translation and dilation of the kernel $\phi(x)$.

$$\phi_{k,l}(x) = 2^{-k/2}\phi(2^{-k}x - l), k, l \in Z \quad (3.11)$$

A series of nested subspaces V_j, which is spanned by the orthonormal basis $\{\phi_{j,k}(x)\}, k \in Z$, forms a multiresolution space. The subspace W_j, which is spanned by the orthonormal basis $\{\psi_{j,k}(x)\}, k \in Z$, is the complementary space of V_j in the subspace V_{j-1}.

$$V_{j-1} = V_j \oplus W_j \quad (3.12)$$

The subspaces V_j and W_j are called the approximate and residue spaces respectively at resolution j.

For the discrete case, the coefficients $h(n)$ and $g(n)$ also play an important role because the continuous forms of $\phi(x)$ and $\psi(x)$ can be neglected, and the coefficients can be directly applied to the discrete signal using the following iterations:

$$a_{j+1}(n) = \sum_k a_j(k)h(k-2n)$$
$$r_{j+1}(n) = \sum_k a_j(k)g(k-2n) \quad (3.13)$$

Where $a_j(n)$ and $r_j(n)$ are the coefficients at resolution j.

Thus, for a J –level discrete wavelet decomposition of the given coefficients $a_0(n)$, a series of coefficients, $\{a_J(n), r_J(n), \ldots, r_1(n)\}$, is obtained. Because of the orthonormality of the wavelet transform, the number of coefficients after the decomposition is equivalent to that before decomposition. The synthesis of the signal from the wavelet coefficients obeys the following:

$$a_j(n) = \sum_k a_{j+1}(k)h(k-2n) + \sum_k r_{j+1}(k)g(k-2n) \quad (3.14)$$

If $\tilde{h}(n)$ and $\tilde{g}(n)$ are defined as

$$\tilde{h}(n) = h(-n), \quad \tilde{g}(n) = g(-n), \quad (3.15)$$

and regarded as the respective impulse responses of the *quadrature mirror filters* (QMF), H_Q and G_Q, which correspond to the halfband lowpass and highpass filter respectively, Eq. (2.24) can be conveniently implemented using convolution and downsampling (here the downsampling rate is 2 which implies that every other sample is omitted) techniques in the signal processing literature. Refer to Fig. 2.1 below for the diagram of the implementation for the 2-D case. If the wavelet decomposition is recursively applied to the output of the lowpass filter, a pyramid-structured decomposition is obtained.

It is interesting to note that the wavelet framework also links the tree-structured wavelet transform with wavelet packets. The library of wavelet packet basis functions $\{W_n\}_{n=0}^{\infty}$ can be obtained from a given W_0 as follows:

$$W_{2n} = \sqrt{2}\sum_k h(k)W_n(2x-k)$$
$$W_{2n+1} = \sqrt{2}\sum_k g(k)W_n(2x-k) \quad (3.16)$$

Where the functions W_0 and W_1 are set to the scaling function $\phi(x)$ and the mother wavelet function $\psi(x)$, respectively. Then, the functions $W_n(2^k x - l), k, l \in Z, n \in N$ form the orthogonal wavelet packet basis. The implementation of the wavelet packets will naturally lead to a tree-structured decomposition, thereby implying that both the outputs of the lowpass and highpass filters are recursively decomposed.

a) Munich and Cambridge morlet filter bank

Morlet wavelet is chosen as the mother wavelet to decompose the audible spectrum for the cochlear filters bank. We use the Morlet wavelet based on the fact

that it has good properties in joint time frequency localization and has a well defined impulse response. In the time domain, the mother wavelet is a high-frequency vibration whose amplitude decreases when the time varies from zero to infinity. The Morlet wavelet transform has proved to be beneficial for many types of wave signal processing and has shown a good performance in tasks like audio coding. The Morlet wavelet shape is not simply analytical, but also very resolute in time and frequency, regular, locally periodic and with non compact support. It is formed by complex values of the shape of a complex sinus modulated by a Gaussian envelope. It is characterized by narrow frequency response, which offers a higher spectral resolution. This wavelet is defined.

IV. Evaluation of Sound Quality

Approval and intelligibility are the two most significant criteria to consider subjective quality. For this purpose, we invited several subjects to hear some mp3 files resulting from the coder based on FFT analysis, then on Munich and Cambridge Morlet wavelet transforms.

a) Simulation Results

In order to evaluate Morlet Munich and Cambridge wavelet compression, we used for various bitrates between 64 kbits/s and 160 Kbits/s some types of sound such as Soul, Slow, Rock, Arabic music and voice (this type of sound contains a dialogue between two persons). The evaluation is based on the compression ratio defined as the quotient between the size of the original file and MP3 file. Tables 2 and 3 contains the type of the sound files used for the test, the average of compression ratio (Comp), their capacity (Cp), their duration (t) and the compression ratio calculated for each bitrate. The weakest compression ratio is given for the flow of 160 Kbits/s and the highest is given for the flow of 64 Kbits/s. However, the weaker the bitrate, the higher the compression ratio is and the less intelligible its quality becomes. Based on the results obtained in tables 2 and 3, the Morlet Munich filters bank takes account of the masking phenomenon and takes account of the critical bands. The second part of our evaluation will focus on comparing the compression ratio obtained from the coder based on wavelet analysis and the coder based on an FFT analysis versus using a bitrate of 128 kbits/s which gives the best compromise between compression ratio and sound quality [10]. Table 1 contains the type of the sound files used for the test, the average compression ratio (Comp) and the compression ratio values using FFT and wavelet (Munich and Cambridge). It reveals that sound compression using Morlet Munich wavelet is the best one. In fact, its average compression ratio presents an improvement of 12.67% and 3.2% in comparison to the FFT and Cambridge coders. The protocol of evaluation consists in listening to the compressed sound file. Then, the possibility is offered to the listeners to listen to it as long as they wish. The listeners are 24: 12 men and 12 women between 15 and 35 years old. Our aim is to classify the three coders for each type of sound in a statistic card as shown in Figure 7. The listeners rate the speech they have just heard on a five-point opinion scale, ranging from "bad" to "excellent". The ratings are assigned integer scores ranging from 1 for"bad" to 5 for "excellent". This comparison, based on results obtained from the statistic card, shows that Morlet Munich coder gives satisfactory results especially for Slow, Soul and Arabic sound. We noticed that for those sound files our coder received a significant number of notes ranging between 3 and 4. Consequently; it gives the best statistics for quality in comparison to the FFT and Cambridge coders.

V. Synthesis Result

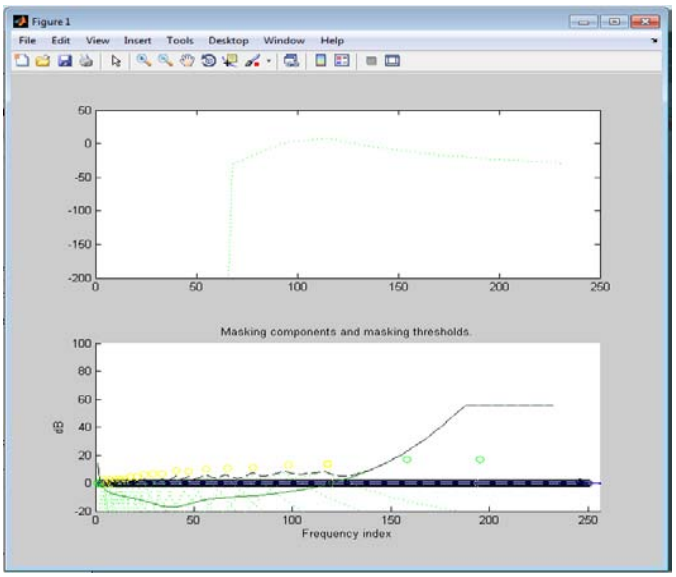


Figure 5.1 : Masking components thresholds

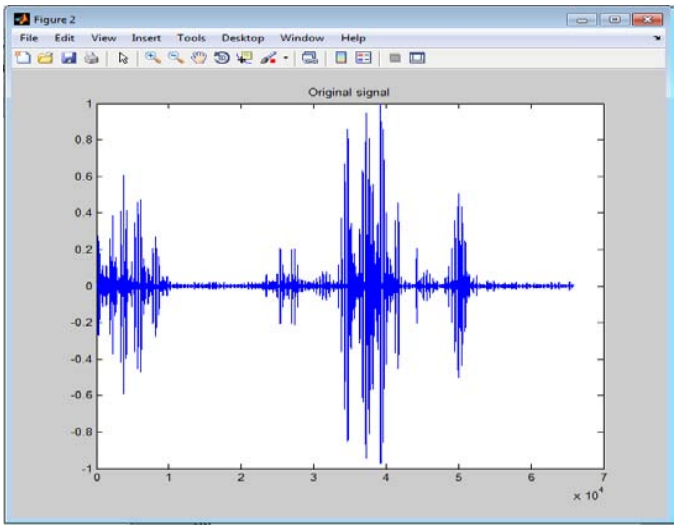


Figure 5.2 : Original Signal

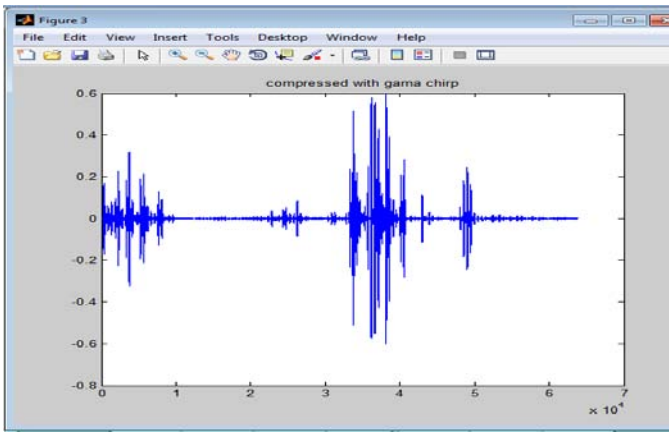


Figure 5.3 : Compressed with game chirp

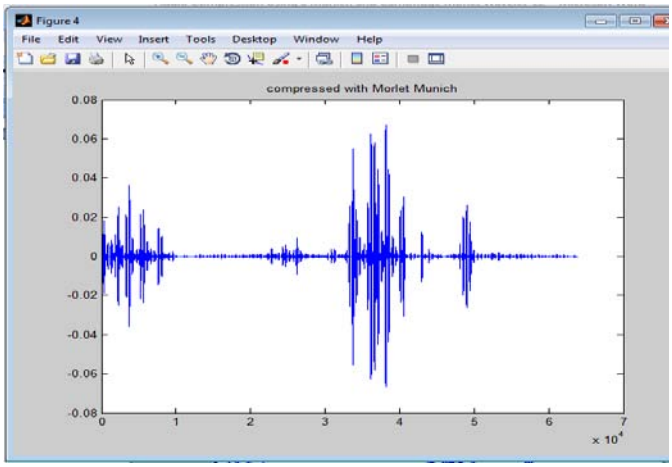


Figure 5.4 : Compressed with Morlet Munich

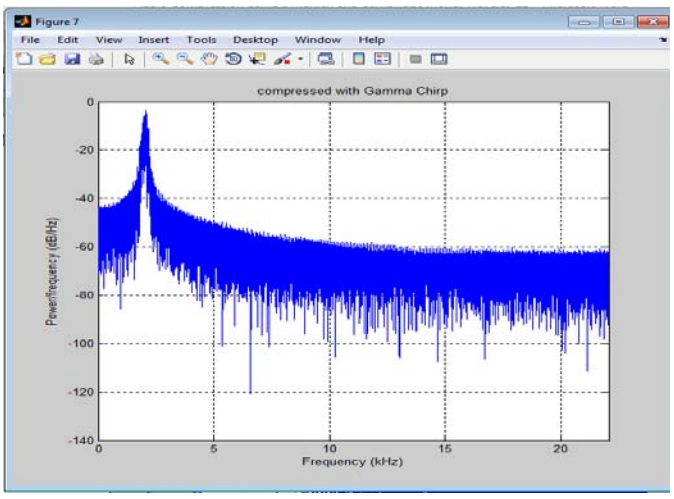


Figure 5.5 : Compressed with game chirp

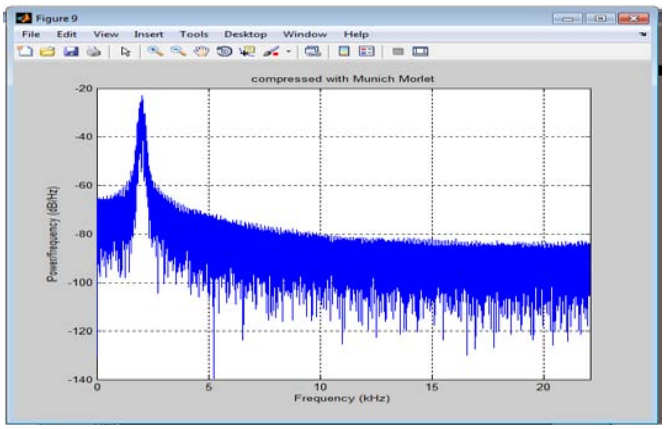


Figure 5.6 : Compressed with Munich Morlet

VI. Conclusion

This paper studies audio compression using a Munich and Cambridge Morlet wavelet transforms. Our goal is to identify the best model that can be applied to the audio coders in sub-bands approximating the critical bands and finally to improve the psycho-acoustic model. This study shows that the Munich Morlet wavelet packet decomposition offers a good performance. In fact, it gives the best compression ratio and sound quality in comparison to the FFT and Cambridge coders. A high selectivity was noticed and can lead to some interesting perspectives on audio coding using this type of psycho-acoustic model.

VII. Acknowledgement

The work is carried out through the research facility at the Department of Computer Science & Engineering, Department of Electronics & Communication Engineering and Department of Electrical & Electronics Engineering, HITS College of Engineering, Bogaram(V), Keesara(M), R.R. District, Hyderabad A.P., and Dept., Electronics & Communication Engineering from VJIT-Hyderabad. The Authors also would like to thank the authorities of JNT Kakinada, SKIT for encouraging this research work.

References Références Referencias

1. ISO /IEC 11172-3 (F), "Information technology - Coding of moving picture and associated audio for digital storage media at up to about 1.5Mbits/s Part3: Audio," 1999.
2. G. Hernandez, M. Mendoza, B. Reusch, and L. Salinas, "Shiftability and filter bank design using Morlet wavelet," Computer Science Society, Nov. 2004, pp. 141- 148.
3. D. Sinha and A. Tewfik,"Low Bitrate Transparent Audio Compression using adapted wavelets," Proc. IEEE Trans.Sig, Dec. 1993, pp. 3463-3479.
4. B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized

fast wavelet packet transform algorithm," Proc. IEEE Transaction on speech and audio processing, vol 47, NO 6, June 1999.
5. Daubechies, "Ten lectures on wavelet," SIAM, Philadelphia, PA, 1992.
6. J. O. SmithIII and J. S. Abel,"Bark and ERB Bilinear transforms," Proc. IEEE Transactions on Speech and Audio Processing, 1999.
7. W. W. Chang and C. T. Wang, "Auditory patterns," Proc. IEEE Transactions on speech and audio processing," vol. 4, NO 2, March 1996.
8. E. Zwicker and E. Terhardt, "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am, vol. 68, NO 5, Nov. 1980.
9. C. Wang and Y. C. Tong,"An improved critical-band transform processor for speech applications," Circuits and Systems, May 2004, pp. 461-464.
10. G. Davidson, L. Fielder, and M. Antill, "High-quality audio transform coding at 128 kbits/s," Proc. IEEE ICASSP, vol.2, 1990, pp. 1117-1120.
11. F. Baumgarte, "A computationally efficient cochlear filter bank for perceptual audio coding storage," Proc. IEEE ICASSP, vol.5, 2001, pp.3265-3268.
12. H. Fletcher, "Auditory patterns," Rev. Mod. Phys, vol.12, 1940, pp. 47-65.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 5 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Efficient Distributed Algorithm using Association Rule Mining for Large Database

By N.K. Sharma & Dr. R.C. Jain
S.A.T.I. Engineering College, India

Abstract - Day by day data is increasing due to effectively computerization, implementation, and digitization in various sectors i.e. science, research, industry, business and many other areas. Data mining is the process of extracting valuable and useful information from this very large database.

Association rule is a concept in which buyer usually by a specific combination of different products together while association rule mining is an important technique to show relationship between various items stored in the database. In this paper we take one master tree called root and various branches process the database rather than a multiple FP-tree.

Keywords : knowledge discovery, FP-tree, a-priori, association rule(s), parallel processing and frequent item set.

GJCST-C Classification : H.2.8

Strictly as per the compliance and regulations of:

Efficient Distributed Algorithm using Association Rule Mining for Large Database

N.K. Sharma [α] & Dr. R.C. Jain [σ]

Abstract - Day by day data is increasing due to effectively computerization, implementation, and digitization in various sectors i.e. science, research, industry, business and many other areas. Data mining is the process of extracting valuable and useful information from this very large database.

Association rule is a concept in which buyer usually by a specific combination of different products together while association rule mining is an important technique to show relationship between various items stored in the database. In this paper we take one master tree called root and various branches process the database rather than a multiple FP-tree.

Keywords : knowledge discovery, FP-tree, a-priori, association rule(s), parallel processing and frequent item set.

I. Introduction

Utilization of information technology in various sectors, the database is increasing day by day. Computerization of land records, registration of vehicles, a collection of various taxes and conduction of any competitive examination etc. require certain attention. Research has been performed on serial data mining but it is not suitable for huge data. In this situation parallel data mining is the viable solution [1] [2] [3] [4]. In parallel data mining available work is to be divided in various processors to get a solution as fast as possible. There are three components to the work i.e. computation, access to the data set and communication between the processor. .Access to the data set is most costly, followed by communication, with computation being relatively cheap [5] [7] [8].

There are three basic strategies [16] [17] [18] for parallelization:

1. Each processor has access to the whole data set, but each processors place in a different part of the search space, starting from a randomly chosen initial position and this strategy is termed as independent search.
2. Each processor restricts itself to generate a particular subset of the set of possible concepts. It completes in two stages. On the first stage, each processor generates complete concepts, but with restriction on the variable values in the same position. (Examine only a subset of the rows of the data set) In the second stage, each processor generates partial concepts but the variables can take any values (examine a subset of the columns) and this strategy is termed as parallelized sequential data mining algorithm.
3. Each processor makes on a partition of the data set (rows) and executes a sequential algorithm. So it builds entire concepts locally not globally and this strategy is termed as replicated sequential data mining algorithms.

Parallel Algorithms provide scalability to large data sets and hence improve response time. Its execution time depends on input size and also on the architecture of the parallel computer. All the algorithms proposed for parallel in shared nothing architecture can also be implemented for distributed architecture because many large databases are distributed in nature.

Apriori algorithms generate candidate item sets and scan databases as many times as the length of the longest frequent item sets whereas in FP-tree algorithms database scan only twice.

a) Existing Strategies

Research has been conducted by various researchers while dealing with large database. Out of which parallel system is found as one of a good strategy. Marfuz (2003) introduced two kinds of methods, parallel association rule mining (PARM) and distributed association rule mining (DARM). In, PARM divides a database into several local databases, and uses parallel multiprocessor shared-nothing environment to mine local databases [19]. The processors need to communicate with each other for the global counts. In case of DARM, association rules discovered from the geographically distributed data sets. The main drawback of DARM is the communication cost.

b) Our Strategies

In this paper, huge database divided into several small databases as per number of processors. The processors have access to only their local database. Processors can generate initial P tree local to their partition. The sequential process of database scanning is divided among the 'n' processors. In one scan, all the information about the transactions local to the processor store in their memory in the form of local P tree data structure and local counts of each item can be simultaneously calculated. Items which have the

Author α : Assistant Professor, Government Engineering College Ujjain, M.P., India. E-mail : nksharma070965@gmail.com
Author σ : Director, S.A.T.I. Engineering College Vidisha, M.P., India. E-mail : dr.jain.rc@gmail.com

support count more than the minimum support are to be included in the subsequent generation of FP tree and the support count stored at each processor is its local count. The local counts of each processor are summed up to get the global counts and the processor can simultaneously prune its infrequent items to get the frequent item set. No information lost is possible in generating FP tree as the FP tree generated at each site, if combined together will exactly replicate the global FP tree. In sequential algorithm, the conditional FP tree of each item present in the header table generates one after the other while in parallel mining, total items in the header table says m; can be divided among n total processors in the distributed environment. This division of work should be such that the amount of processing for generating the patterns at each processor is comparable. The division of items among the processors should be in such a way that the processor which gets the item least support count also gets the item with the highest support count. The items with lower support counts should be equally divided among the nodes. The next step is to calculate the global conditional pattern base of the item. If local conditional FP trees of an item are collected from all the processors and send to the destination processor, the global conditional FP tree can be generated. Hence the conditional FP tree can be sent in the form of conditional pattern base from which it is generated.

II. Work Already Done

Count Distribution (CD), Data Distribution (DD), Candidate Distribution, Intelligent data distribution (IDD) and Hybrid distribution (HD) algorithm briefly described [13] [14] [15] below:

a) Count Distribution (CD)

In which all processors generate the entire candidate set from Lk-1. Each processor can thus independently get partial support of the candidates from its local database partition and sum up to get global counts by exchanging local counts with other processors. Once global frequent sets have been determined, next candidate item sets can be determined in parallel at all processors. The focus is on minimizing communication. It does so even at the expense of carrying redundant computations in parallel. The aggregate memory of the system is not exploited effectively.

b) Data Distribution (DD)

Uses the total system memory by generating disjoint candidate sets on each processor. However, to generate the global support, each processor must scan the entire database in all iterations. In DD algorithm Contention is a major problem due to this processor may remain idle at the time of communication.

c) Candidate Distribution

Algorithm partitions the candidates during iteration l, so that each processor can generate disjoint candidates independent of other processors. The partitioning uses heuristics based on support, so that each processor gets an equal amount of work. The choice of the redistribution pass involves a tradeoff between decoupling processor dependence as soon as possible and waiting until sufficient load balance can be achieved. No local data send, only global values exchanged and data are received asynchronously and the processors do not wait for the complete pruning information to arrive from all the processors but repartitioning is expensive.

d) Intelligent Data Distribution (IDD)

The locally stored portions of the DB can be sent all the other PEs by using the linear-time ring-based all-to-all broadcast. Although DD divides the candidates equally among the processors, it fails to divide the work done on each transaction. IDD algorithms switch to CD once the candidates fit in the memory. Instead of a round-robin candidate partitioning, IDD performs a single-item, prefix based partitioning. Pseudo code for data movements using sending Buffer (SBuf) and receiving Buffer (RBuf).

Hybrid distribution (HD) algorithm combines CD and IDD. It partitions the P processors into G equal – sized groups, where each group is considered a super processor. Count Distribution is used among the G super processors, while the P/G processors in a group use Intelligent Data Distribution. The database is horizontally partitioned among the G super processors, and the candidates are partitioned among the P/G processors in a group.

The advantages of this algorithm are:

i. Provide good load balance.
ii. Handle much larger databases as compared to CD.
iii. Provide enough computation work by maintaining a minimum number of Candidates per PE cut down the amount of data movement to 1/G of the IDD.
iv. Keep processors busy, especially during later iterations.

e) FP-Growth Algorithm

FP stands for Frequent Patterns. This algorithm [6] makes use of an FP-tree structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns. FP-Growth is an efficient method for mining complete set of frequent patterns by pattern fragment growth. Efficiency of mining is achieved with three techniques. First, a large database is compressed into a highly condensed, much smaller data structure, which avoids costly repeated database scans. Second, FP tree-based mining adopts a pattern fragment growth method to avoid the costly generation of a large number of

candidate sets. And third, a partitioning based, divide and conquer method is used to decompose mining task into a set of smaller tasks for mining confined patterns in conditional databases, which dramatically reduces the search space. FP-Growth is efficient for mining both long and short frequent patterns.

It avoids the costly candidate generation and it has better performance and efficiency than Apriori like algorithms but it takes two complete scans of the database and it uses a recursive routine to mine patterns from a conditional pattern base.

f) P-Tree Algorithm

A Pattern Tree [9] [10], unlike FP Tree, which contains the frequent items only, contains all the items that appear in the original database. We can obtain a P-tree through one scan of database and get the corresponding FP-tree from the P-tree later. An FP tree is a sub-tree of the P-tree with a specified support threshold, which contains those frequent items that meet this threshold and hereby excludes infrequent items. We do this by checking the frequency of each node along the path from root to leaves. It uses the same mining process as used by FP-Growth algorithm [6].

It scans the original database only once and in case support threshold changes, we need not re-scan the database but it uses a recursive mining process.

g) Inverted Matrix Algorithm

This association rule-mining algorithm is based on the conditional pattern concept [6]. The algorithm [11] [12] is divided into two main phases. The first one, considered pre-processing, requires two full I/O scans of the dataset and generates a data structure called Inverted Matrix. In the second phase, the Inverted Matrix is mined using different support levels to generate association rules. The mining process might take in some cases less than one-full I/O scan of the data structure in which only frequent items based on the support given by the user are scanned and participate in generating the frequent patterns. The Inverted Matrix layout combines the horizontal and vertical layouts with the purpose of making use of the best of the two approaches and reducing their drawbacks as much as possible. The idea of this approach is to associate each item with all transactions in which it occurs (i.e. An inverted index), and to associate each transaction with all its items using pointers.

For computing frequencies, it relies first on reading sub-transactions for frequent items directly from the Inverted Matrix [6]. Then it builds independent relatively small trees for each frequent item in the transactional database. Each such tree is mined separately as soon as they are built, with minimizing the candidacy generation and without building conditional sub-trees recursively.

It uses a simple and non-recursive association rule mining process and the inverted matrix can be made disk resident, so it performs well for large data sets but it makes two scans of the original database and the complexity of developing an inverted matrix is a bit high.

III. Proposed Methodology

a) Mining Process

To find out frequent itemsets having minimum support and to generate strong rules having minimum confidence, Apriori algorithm is a bottom – up search based algorithm in which any subset of large item set must also be large.

b) Proposed Assumptions

- Horizontally partitioned database.
- Sites communication through message passing which reduce the number of messages passed and confine substantial amount of processing at local sites.

IV. Proposed Algorithm

- Select database and then arrange it in required proper format.
- Generate the initial P-Tree and find out the local counts of the items.
- Generate global counts of items by exchanging the local counts at each site.
- Generate a FP - Tree at each processor using its local data
- Distribute the frequent items such that comparison of computation required at each site is comparable
- Generate the pattern base for each item and start sending them to the corresponding processor
- Generate conditional FP-Trees using the data received from all the sites and those generated locally
- At the allocated processor mine frequent patterns for the corresponding items.

V. Implementation Example Database

To expound the effectiveness of the approach specified above a Student counseling data set (SCDS) has been considered, in which a separate agency has conducted an examination and on the basis of merit candidates have filled online choices of institutions and branches. The total number of combination of the choices was around 1200 and the total filled choices for institution were around 25 lakhs. To process such large data sequentially, it will roughly take 10 to 15 minutes. Hence for quick response it becomes necessary to use distributed architecture. We have taken Student counseling data set (SCDS), with three attributes, namely RollNo, Choice_Sequence and Branch_ID. Similarly Branch_ID consists of branch ID and college

name. Following Tables show Student Admission Data Set (SADS). Relational view of the same is being illustrated below.

Student Choice Data		
Roll No	Student Choice_ Sequence	Branch_ ID
1001	1	4
1002	2	2
1003	3	6
1004	4	11
1005	5	7
1006	6	9

Table 1 : Student Choice Data

Branch ID Generation		
Branch_ ID	College_Name	Branch
1	C1	CS
2	C1	IT
3	C1	EC
4	C2	CS
5	C2	IT

Table 2 : Branch ID Generation

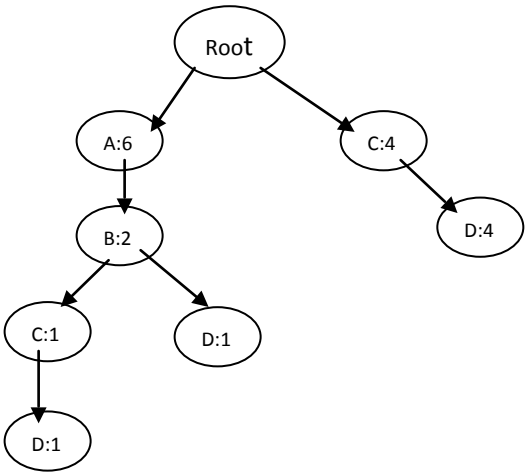
Above specified data is then treated with the proposed approach and transactional data set is obtained for the above given data which is expounded in Table 3.

Transactional Data Set				
Transaction_ID	A	B	C	D
T1	1	1	1	1
T2	1	0	1	1
T3	1	1	0	1
T4	1	0	1	1
T5	1	0	1	1
T6	1	0	1	1

Table 3 : Transactional Data Set for Student Choice Data of Table 1

VI. Tree Generated

On the basis of data mentioned in table 3, a tree is generated by the software



VII. Rules Generated and Result Analysis

a) Rules Generated

Rules are generated by taking minimum support is 75% and minimum confidence is 80%. The generated rules are shown in Table 4.

Rule	Confidence
C1 -> C3	83.33333
C3 -> C1	100
C1 -> C4	100
C4 -> C1	100
C3 -> C4	100
C4 -> C3	83.33333
C1, C3 -> C4	100
C4 -> C1, C3	83.33333
C1 -> C3 ,C4	83.33333
C3, C4 -> C1	100

Table 4 : Rules Generated

b) Result Analysis

Depending on the number of transactions the run time is calculated using Sequential FP–Tree algorithm as well as by using the proposed distributed approach and the comparison is expounded in Table 5.

S.No.	No. of Transactions	Run Time using Seq. FP -Tree (in seconds)	Run Time using Proposed Distributed approach (in seconds)
1	1000	5	6
2	5000	9	8
3	7500	16	11
4	10000	23	15

Table 5 : Result Analysis

Scale up shows that the proposed algorithms handles larger problem sets when more processors are available. Scales up experiments were performed where the size of the database was increased in direct proportion to the number of nodes in the system. The speedup gives decrease in the response time with the increase of number of processors. All the experiments are performed on a Xeon 2.8 GHz machine with 4 GB RAM running on Windows 7 platform. All the programs are written in JAVA platform.

VIII. Conclusion

The main focus of the paper is to design a mining algorithm for distributed environment using just a single database scan and distributing the computation work equally within the processors. The proposed algorithm allows efficient mining of patterns as 'n' items are considered for finding patterns simultaneously, where n is the number of processors. Computation time is reduced as the workload is divided amongst various processors. A processor receives the patterns from each processors one after the other and if one machine is slow other machines are also effected. Hence the process should be improved by the simultaneous construction of conditional pattern tree along with the improvement in the receive operation.

Acknowledgements

Our sincere thanks to our colleague Mr. Manoj Yadav, Software Consultant in Bhopal, Madhya Pradesh, India for the immense support you have provided us for publishing this paper.

References Références Referencias

1. R. Agrawal, T. Imienski and A. Swamy, "Database Mining: A Performance Perspective, IEEE Tran. On Knowledge and Data Engg." December, 1991.
2. R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 1993.
3. Margaret H. Dunham, Yongqiao Xiao, Southern Methodist University, Dallas, Texas and Le Gruenwald, Zahid Hossain, University of Oklahoma, Norman, UK, "A Survey of Association Rules".
4. Jong Soo Park, Ming-Syan Chenand Philip S. Yu, "An effective hash-based algorithm for mining association rules," In Proceedings of 1995 ACM-SiGMOID international Conference on Management of Data, 1995.
5. Mohammed J. Zaki, Rensselaer Polytechnic Institute, "Association Mining: A Survey", IEEE Concurrency, 1999.
6. Jiawei Han, Jian Pei, and Yiwen Yin, "Mining frequent patterns without candidate generation", Technical Report CMPT99-12, School of Computing Science, Simon Fraser University, 1999.
7. Mohammad El-Hajj and Osmar R. Zaiane, "Parallel Association Rule Mining with Minimum Inter-process Communication", In Proc. of IEEE Int. Conf. On Database and Expert Systems Applications, 2003.
8. R. Agarwal and J. Shafer, "Parallel Mining of Association Rules," IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, December 1996.
9. D. W. Cheung, V.T. Ng, A.W. Fu and Y Fu, "Efficient Mining of Association Rules in Distributed Databases", IEEE Tran. On Knowledge and Data Engg. December, 96.
10. A.Y. Zomya, T.E. Ghazawi and O. Frieder, "Parallel and Distributed Computing for Data Mining", IEEE Concurrency, Oct./Nov. 1999.
11. Skillicorn, "Strategies for Parallel Data Mining", IEEE Concurrency, Nov. 1999.
12. A. Mueller, "Fast and Sequential Algorithms for Association Rule Mining." A comparison, Tech Report CS-TR-3515, Univ. of Maryland, College Park. Md. 1995.
13. Margaret H. Dunham and Yongqiao Xiao, Southern Methodist University, Dallas, Texas and Le Gruenwald, Zahid Hossain, University of Oklahoma, Norman UK, " A survey of Association Rules."
14. Mohammed J. Zaki, Rensselaer Polytechnic Institute, "Parallel and Distributed Association Mining: A Survey", IEEE Concurrency 1999.
15. E.H. Han, G. Karypis, and V. Kumar, "Scalable Parallel Data Mining for Association Rules," IEEE Transactions on Knowledge and Data Engineering, Vol. 12, No. 3, May/June 2000.
16. T. Shimomura and S. Shibusawa, "Performance Evaluation of Distributed Algorithms for Mining Association Rules on Workstation Cluster", IEEE 2000.
17. Hao Huang, Xindong Wu and Richard Relue, "Association Analysis with One Scan of Databases", IEEE 2002.
18. O.R. Zaiane, M.E. Hajj, "Parallel Association Rule Mining with Minimum Inter-Processor Communication", 2003 IEEE, 14th International Workshop on Database and Expert Systems Applications.
19. Fan Wu, Ya-Han Hu, Tz Ke Wu, "A Novel and Efficient Distributed Data Mining Algorithm Based on Frequent Pattern-Tree", Int'l Conf. Data Mining | DMIN'09 |.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 5 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Review of Multilingual Glossing and Related Formats

By Preeti Kamboj & Punita Meelu
Doon Valley Institute of Engineering & Technology, Karnal

Abstract - This review paper gives a Historical review of language documentation formats and various glossing techniques available. There is a need of a Multilingual glossing format that is capable of providing features described in this paper. The paper introduces advanced glossing and tools available for the same.

Keywords : *linguistics, annotations, glossing, glossing tables, at, documentation formats.*

GJCST-C Classification *: D.0*

Strictly as per the compliance and regulations of:

A Review of Multilingual Glossing and Related Formats

Preeti Kamboj [α] & Punita Meelu [σ]

Abstract - This review paper gives a Historical review of language documentation formats and various glossing techniques available. There is a need of a Multilingual glossing format that is capable of providing features described in this paper. The paper introduces advanced glossing and tools available for the same.

Keywords : linguistics, annotations, glossing, glossing tables, at, documentation formats.

I. Introduction

In recent years, the documentation of languages, especially of endangered languages, has received a growing interest within the linguistic community. Most researchers agree that the core of language documentation should consist primarily of recorded and transcribed texts which should not only be translated but also annotated or glossed to be of use for a wide range of purposes.

However, although Lehmann proposes as the principal aim of that format "to make the grammatical structure [of a text in an unknown language] transparent", all it provides is a rendering of the "meaning or function" of the individual morphemes.

A type of format widely used, especially in the context of grammar writing within a functional framework, is interlinear morphemic translations, or interlinear glossing (IG) for short, first systematized by C. Lehmann.

This has proven useful when exemplifying established facts or illustrating a discussion, but it is by no means sufficient for the aims of language documentation. These aims include that it should be possible to write a grammar of the language or variety being documented, given a sufficient large number of texts that are completely documented [1].

a) Types of Glossing Information

1. *Syntactical:* Syntactical information includes Syntax related information of the language.
2. *Morphological:* Morphological information includes various lines that contain the instances, phonological, semantic, categorical, structural and relational information.

Author α : Student, M.Tech(C.S.E), Doon Valley Institute of Engineering & Technology, Karnal.
Author σ : Assistant Professor C.S.E Department, Doon Valley Institute of Engineering & Technology, Karnal.

AG (Advanced Glossing) is a general format that does not stipulate details of a possible technical implementation. However, it is obvious that in the digital age any such format should be applicable by means of appropriate computer software.

b) Applications of Advanced Glossing

1. To provide transparent structure to any documentation regardless of languages.
2. Multilingual Environment.
3. Feature Extraction
4. Interactive teaching and E-Learning
5. Application in computer aided machine translation system.

II. Requirements for a Language Documentation Format

The following conditions (which may not be independent) as minimal requirements that any language documentation format must meet if the documentation is to be suitable for the purposes of linguistics:

i. It must be possible to write a grammar of the language or variety being documented given a sufficiently large number of texts completely documented within that format.
ii. The language used as the documentation language must be clearly interpretable; in particular, it must be possible to clearly distinguish between phonetic, phonological (phonemic), morphological, syntactic, and semantic information in the documented text.
iii. The documentation format must allow both for partial documentation of a text and for the gradual, systematic filling in of gaps during the documentation process.
iv. The documentation format must be such that information gathering for the complete documentation of a text (which may not be possible in all cases) can, in principle, be achieved under field conditions.

The character of these conditions: they are (i) minimal and (ii) conditions for the purposes of linguistics. Even for these purposes, additional conditions easily come to mind, and the above requirements do not yet cover the conditions imposed on language documentation for purposes outside linguistics.[2]

These requirements must be met, if one wants a language to be documented. The Typological Glossing doesn't meet these requirements. Thus the need for something more is felt. Then came the Advanced Glossing. Advanced Glossing meets these all requirements and along with that provide the structure in the form of Advanced Glossing Tables. (AGTs)

III. Related Work

Sebastian Drude and et. al. in their paper "Advanced Glossing: A Language Documentation Format and its implementation with Shoebox"[1] presents a proposal for a general glossing format designed for language documentation, and a specific setup for the Shoebox-program that implements Advanced Glossing to a large extent. AG provides specific lines for different kinds of annotation – phonetic, phonological, orthographical, prosodic, categorical, structural, relational, and semantic, and it allows for gradual and successive, incomplete, and partial filling in case that some information may be irrelevant, unknown or uncertain. The implementation of AG in Shoebox sets up several databases. Each documented text is represented as a file of syntactic glossings. The morphological glossings are kept in a separate database.

Hans-Heinrich Lieb, Sebastin Drude presents the requirements for a language documentation format in "Advanced Glossing: A language Documentation Format (Working Paper, November 2000)" [2]. The language document format must be suitable for the language linguistics format. It shows that the Typological Glossing doesn't meet these requirements. They also emphasizes on the need of the presentation of the Advances Glossing.

J D Riding presents their work in "Statistical Glossing - Language Independent Analysis in Bible Translation" [3]. Bible translation sets a number of particular challenges for machine translation and analysis. The nature of the work is such that translators are often working with local vernacular languages for which there are little in the way of lexica and other linguistic databases. Commercial text processing and translation systems are rarely able to contribute. Translation Consultants (TCs) charged with advising translation teams often have few computer based aids to assist them in their work. This paper describes the development of the Statistical Glossing Tool (SGT) which ships with the Paratext translation editor. SGT is a language independent method for the analysis of bible translations. It provides an objective assessment which TCs can use to help them identify key issues in a new translation where further work may be needed. SGT requires no information about target languages other than the text itself.

Puneet Kamboj, Jatin Aneja, Ajay Malhotra provides their work in "Web and Mobile based Dynamic Glossing tool for Foreign Language Text". [4] In that they says that Annotation is add-on information attached at a particular point in a document. Much software exists for attaching annotations to document text. This thesis uses a concept of hot words similar to annotations for attaching information to keywords or phrases in published books. Hot words provides a new way of describing keywords or phrases in books by attaching text, image, audio and video based information to them. To ensure global reach this software is presented on web and mobile based platforms supporting multiple languages with an ability to add more languages.

The software developed serves as a great tool for e-learning by providing more information on keywords or phrases. Students can also contribute by suggesting additional annotations. This keeps the information ever expanding.

The Gloss Tool is based on three tier architecture with web service used to develop business layer to expose complete API over the web.

IV. Conclusion

There exist various methods and tools to incorporate multilingual information along with annotations. One of such tools is SIL Toolbox.

This paper describes techniques and Applications related to incorporating Glossing Information into a Language independent format. The format must support Advanced glossing Tables, hence ability for future extension.

The paper also depicts work carried out in the field of Advanced Glossing, majorly by "Sebastian Drude".

V. Future Scope

Although there exists a lot of work on Language Glossing and Advanced Glossing Tables. I was looking for a method to actually extract information from a Multimedia Glossed Table, regardless of Language boundaries.

A feature Extraction tool or Method would help many endangered Languages in following ways:

- It can be used to Extract Dictionaries.
- It can be used to Extract Morphological Information.
- It can be used to extract the syntactic information.
- It can be used to translate most important works in that Language.
- Can be applied in computer aided Statistical machine translation systems.

References Références Referencias

1. Sebastian Drude “Advanced Glossing - a language documentation format and its implementation with Shoebox.
2. Hans-Heinrich Lieb, Sebastin Drude, "Advanced Glossing: A language Documentation Format"(Working paper, November 2000).
3. J D Riding “Statistical Glossing - Language Independent Analysis in Bible Translation”.
4. Puneet Kamboj, Jatin Aneja, Ajay Malhotra “Web and Mobile based Dynamic Glossing tool for Foreign Language Text”.
5. Integrated Data Management and Analysis for the Field Linguistic, The Linguist’s Shoebox: Tutorial and User’s Guide.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 5 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Review Paper on Clustering Techniques

By Amandeep Kaur Mann & Navneet Kaur
RIMT, Mandi Gobindgarh PTU

Abstract - The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Clustering is a significant task in data analysis and data mining applications. It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Data mining can do by passing through various phases. Mining can be done by using supervised and unsupervised learning. The clustering is unsupervised learning. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. Grid Density based algorithm uses the multiresolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time. In this survey paper, an analysis of clustering and its different techniques in data mining is done.

Keywords : *data mining, clustering, classification of clustering, supervised, unsupervised.*

GJCST-C Classification : *H.3.3*

Strictly as per the compliance and regulations of:

Review Paper on Clustering Techniques

Amandeep Kaur Mann [α] & Navneet Kaur [σ]

Abstract - The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Clustering is a significant task in data analysis and data mining applications. It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Data mining can do by passing through various phases. Mining can be done by using supervised and unsupervised learning. The clustering is unsupervised learning. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. Grid Density based algorithm uses the multiresolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time. In this survey paper, an analysis of clustering and its different techniques in data mining is done.

Keywords : data mining, clustering, classification of clustering, supervised, unsupervised.

I. Introduction

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Data mining is also known as the analysis step of the knowledge discovery in databases (KDD). Knowledge discovery means to "develop something new". Data mining practice has the four main everyday jobs. These are Anomaly detection, Association, Classification, Clustering. Anomaly detection is the recognition of odd data records, that may be remarkable or data errors that involve further investigation. Association rule learning is the process to find the relationships between the variables. In this, relations are set up between the variables to create the new information that is needed for some purpose. Classification is the assignment of generalizing the known structure to apply to new data like in an e-mail process might attempt to categorize an e-mail as "legitimate" or as "spam". Clustering is a significant task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Cluster is an ordered list of data which have the familiar characteristics. Data mining is a multi-step process. In data mining data can be mined by passing through various phases.

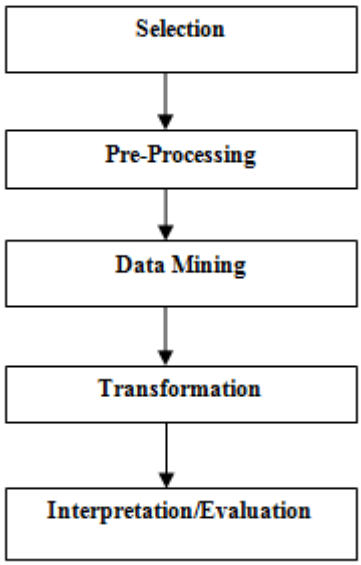


Figure 1 : Phases of Data Mining

In Data Mining the two types of learning sets are used, they are supervised learning and unsupervised learning.

a) Supervised Learning

In supervised training, data includes together the input and the preferred results. It is the rapid and perfect technique. The accurate results are recognized and are given in inputs to the model through the learning procedure. Supervised models are neural network, Multilayer Perceptron and Decision trees.

b) Unsupervised Learning

The unsupervised model is not provided with the accurate results during the training. This can be used to cluster the input information in classes on the basis of their statistical properties only. Unsupervised models are for dissimilar types of clustering, distances and normalization, k-means, self organizing maps.

II. Clustering

Clustering is a major task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups. Cluster is an ordered list of data which have the familiar characteristics. Cluster analysis

can be done by finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. Clustering is an unsupervised learning process. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. The superiority of a clustering result depends on equally the similarity measure used by the method and its implementation. The superiority of a clustering technique is also calculated by its ability to find out some or all of the hidden patterns. Similarity of a cluster can be expressed by the distance function. In data mining, there are some requirements for clustering the data. These requirements are Scalability, Ability to deal with different types of attributes, Ability to handle dynamic data, Discovery of clusters with arbitrary shape, Minimal requirements for domain knowledge to determine input parameters, Able to deal with noise and outliers, Insensitive to order of input records, High dimensionality, Incorporation of user-specified constraints, Interpretability and usability. The types of data that are used for analysis of clustering are Interval-scaled variables, Binary variables, Nominal, ordinal, and ratio variables, Variables of mixed types [1]. The five types of clusters are used in clustering. The clusters are divided into these types according to their characteristics. The types of clusters are Well-separated clusters, Center-based clusters Contiguous clusters, Density-based clusters and Shared Property or Conceptual Clusters. Many applications of clustering are characterized by high dimensional data where each object is described by hundreds or thousands of attributes. Typical examples of high dimensional data can be found in the areas of computer vision applications, pattern recognition, and molecular biology [8]. The challenge in high dimensional is the curse of dimensionality faced by high dimensional data clustering algorithms, basically means the distance measures become gradually more worthless as the number of dimensions increases in the data set. Clustering has an extensive and prosperous record in a range of scientific fields in the vein of image segmentation, information retrieval and web data mining.

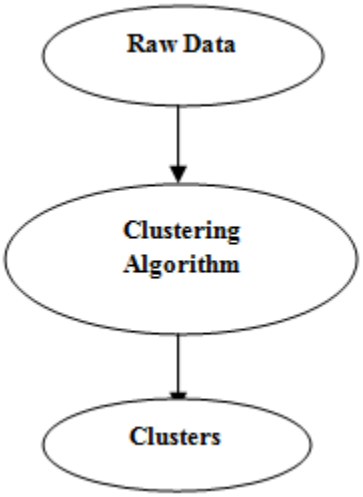


Figure 2 : Stages of Clustering

III. Classification of Clustering

Clustering algorithms can be categorized into partition-based algorithms hierarchical-based algorithms, density-based algorithms and grid-based algorithms. These methods vary in (i) the procedures used for measuring the similarity (within and between clusters) (ii) the use of thresholds in constructing clusters (iii) the manner of clustering, that is, whether they allow objects to belong to strictly to one cluster or can belong to more clusters in different degrees and the structure of the algorithm. Irrespective of the method used, the resulting cluster structure is used as a result in itself, for inspection by a user, or to support retrieval of objects [5].

a) Partitioning Algorithms

Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective function. One such criterion functions is minimizing square error criterion which is computed as,

$$E = \sum \sum || p - m_i ||^2$$

Where p is the point in a cluster and m_i is the mean of the cluster. The cluster should exhibit two properties, they are (a) each group must contain at least one object (b) each object must belong to exactly one group. The main drawback of this algorithm is whenever a point is close to the center of another cluster; it gives poor result due to overlapping of data points [4]. It uses several greedy heuristics schemes of iterative optimization. There are many methods of partitioning clustering; they are k-mean, Bisecting K Means Method, Medoids Method, PAM (Partitioning around Medoids), CLARA (Clustering Large Applications) and the Probabilistic Clustering [8]. For a given *k*, the *k-means* algorithm consists of four steps:

(1) Choose initial centroid at arbitrary.
(2) Allocate each object to the cluster with the adjacent centroid.
(3) Calculate each centroid as the mean of the objects assigned to it.
(4) Reiterate previous 2 steps until no change.

This algorithm is applicable only when *mean* is defined (what about categorical data?). It requires specifying *k,* the number of clusters, in advance which is very difficult. It is not able to handle noisy data and outliers. It is not suitable to discover clusters with non-convex shapes. For handling the categorical data, the algorithm *k-modes* are developed. It replaces the means of clusters with modes. It uses the latest dissimilarity procedures to deal with categorical objects and use a frequency-based method to revise modes of clusters. For a mixture of categorical and numerical data the *k-prototype* is used.

b) Hierarchical Algorithm

Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. This method is based on the connectivity approach based clustering algorithms. It uses the distance matrix criteria for clustering the data. It constructs clusters step by step. Hierarchical clustering generally fall into two types: In hierarchical clustering, in single step, the data are not partitioned into a particular cluster. It takes a series of partitions, which may run from a single cluster containing all objects to 'n' clusters each containing a single object. Hierarchical Clustering is classified as

A. Agglomerative Nesting
B. Divisive Analysis

c) Agglomerative Nesting

It is also known as AGNES. It is bottom-up approach. This method construct the tree of clusters i.e. nodes. The criteria used in this method for clustering the data is min distance, max distance, avg distance, center distance. The steps of this method are:

(1) Initially all the objects are clusters i.e. leaf.
(2) It recursively merges the nodes (clusters) that have the maximum similarity between them.
(3) At the end of the process all the nodes belong to the same cluster i.e. known as the root of the tree structure.

d) Devise Analysis

It is also known as DIANA. It is top-down approach. It is introduced in Kaufmann and Rousseeuw (1990). It is the inverse of the agglomerative method. Starting from the root node (cluster) step by step each node forms the cluster (leaf) on its own. It is implemented in statistical analysis packages, e.g., plus.

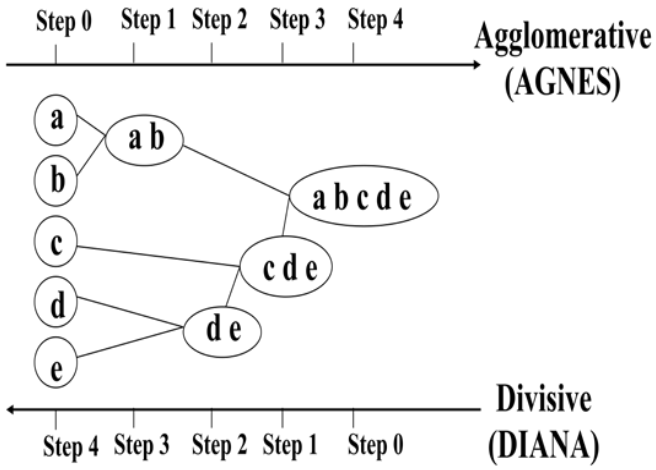


Figure 3 : Representation of AGNES and DIANA

Advantages of hierarchical clustering [2]

(1) Embedded flexibility with regard to the level of granularity.
(2) Ease of handling any forms of similarity or distance.
(3) Applicability to any attributes type.

Disadvantages of hierarchical clustering [2]

(1) Vagueness of termination criteria.
(2) Most hierarchical algorithm does not revisit once constructed clusters with the purpose of improvement.

e) Density Based Algorithms

Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. It is able to find the arbitrary shaped clusters and handle noise. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The density based algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) is commonly known. The Eps and the Minpts are the two parameters of the DBSCAN [6]. The basic idea of DBSCAN algorithm is that a neighborhood around a point of a given radius (ε) must contain at least minimum number of points (MinPts) [6]. The steps of this method are:

(1) Randomly select a point ***t***
(2) Recover all density-reachable points from ***t*** wrt *Eps* and *MinPts*.
(3) Cluster is created, if ***t*** is a core point
(4) If ***t*** is a border point, no points are density-reachable from ***t*** and DBSCAN visits the next point of the database.
(5) Continue the procedure until all of the points have been processed.

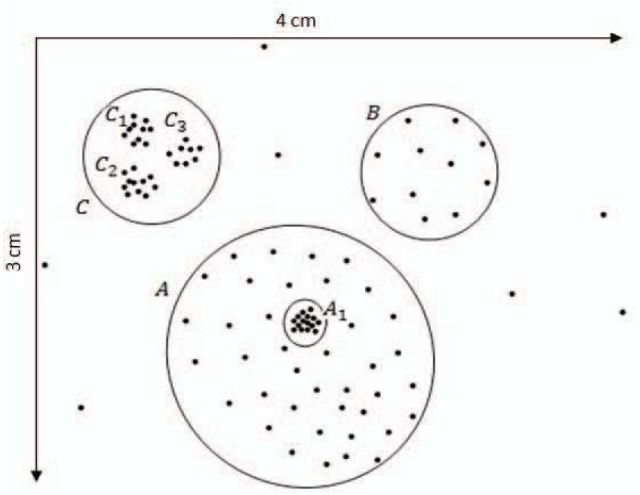


Figure 4 : An example of dataset with different Densities [7]

f) Grid Density Based Algorithms

Grid Density based clustering is concerned with the value space that surrounds the data points not with the data points. This algorithm uses the multiresolution grid data structure and use dense grids to form clusters. It first quantized the original data space into finite number of cells which form the grid structure and then perform all the operations on the quantized space. Grid based clustering maps the infinite amount of data records in data streams to finite numbers of grids. Its main distinctiveness is the fastest processing time, since like data points will fall into similar cell and will be treated as a single point. It makes the algorithm self-governing of the number of data points in the original data set. Grid Density based algorithms require the users to specify a grid size or the density threshold, the problem here arise is that how to choose the grid size or density thresholds. To overcome this problem, a technique of adaptive grids are proposed that automatically determines the size of grids based on the data distribution and does not require the user to specify any parameter like grid size or the density threshold. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE. These methods are efficient only for low dimensions. Among the huge number of cells most are empty and some may be failed with one point. It is impossible to determine the data distribution with such a coarse grid structure. Fine grid size leads to the huge amount of computation, while coarse grid size results the low quality of clusters. The algorithm OPTICS is proposed for the purpose of high dimensional data.

The steps of the grid based algorithm are:

1. Creating the grid structure, in other words divide the data space into a finite number of cells.
2. Calculating the cell density for each cell
3. Sorting of the cells according to their densities.
4. Identifying cluster centers.
5. Traversal of neighbor cells.

There are various algorithms used for clustering the data items into the clusters. Among them the Grid Density algorithms perform well over the time complexity as well as on the high dimensional data.

IV. Conclusion

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Clustering is a significant task in data analysis and data mining applications. It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. Grid algorithm uses the multiresolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time.

Table 1 : Comparison of time complexity among Density and Grid Density algorithm

Algorithm	Review
DBSCAN	Time complexity of algorithm is $O(n)^2$ and not suitable for environments with different densities.
OPTICS	It has the problem of overlapped clusters. Time complexity is also $O(n)^2$.
VDBSCAN	Choosing several epsilons rather than one global epsilon value. Time complexity is O (time complexity of DBSCAN*t), in which t is the number of iterations of algorithm.
LDBSCAN	Using the concepts of *(LOF)* and *(LRD).* It strongly influences the output of clustering and the algorithm has no guidance to select appropriate values.
GMDBSCAN	It works on local density value. And constructs the SP-Tree after dividing into grids. The computational and time complexity both are high
P-DBSCAN	It concentrates on clustering spatial data such as geo-tagged images and GPS. It changes the core point into photo.
MSDBSCAN	The time complexity of the algorithm is still the drawback of the algorithm like DBSCAN.
GDCLU	It generates major clusters by merging dense grids. The time complexity of the algorithm is better than others.

Acknowledgement

The author would like to thanks the Department of Computer Science & Engineering of RIMT Institutes near Floating Restaurant, Sir Hind Side, Mandi Gobindgarh-147301, and Punjab, India.

References Références Referencias

1. J.Daxin, C.Tang and A. hang (2004) Cluster Analysis for Gene Expression Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, Vol. 16, Issue 11, pp. 1370-1386.
2. Pradeep Rai and Shubha Singh (2010) A Survey of Clustering Techniques, International Journal of Computer Applications (0975 – 8887) Vol 7– No.12, pp. 1-5
3. V.Kavitha , M.Punithavalli (2010) Clustering Time Series Data Stream – A Literature Survey, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 1, pp. 289-294.
4. S. Anitha Elavarasi and Dr. J. Akilandeswari (2011) A Survey On Partition Clustering Algorithms, International Journal of Enterprise Computing and Business Systems.
5. S.Vijayalaksmi and M Punithavalli (2012) A Fast Approach to Clustering Datasets using DBSCAN and Applications (0975 – 8887) Vol 60– No.14, pp. 1-7.
6. Cheng-Far Tsai and Tang-Wei Huang (2012) QIDBSCAN: A Quick Density-Based Clustering Technique idea International Symposium on Computer, Consumer and Control, pp. 638-641.
7. Gholamreza Esfandani, Mohsen Sayyadi and Amin Namadchian (2012) GDCLU: a new Grid-Density based Clustering algorithm, 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 102-107.
8. Sunita Jahirabadkar and Parag Kulkarni (2013) Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms, International Journal of Computer Applications (0975 – 8887) Vol 63– No.20, pp. 29-35.
9. Guojun Gan, Chaoqun Ma, Jianhong Wu, *Data Clustering: Theory, Algorithms, and Applications.*
10. Pavel Berkhin, *Survey of Clustering Data Mining Techniques*, Accrue Software, Inc.

This page is intentionally left blank

Global Journals Inc. (US) Guidelines Handbook 2013

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

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

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

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

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

AUXILIARY MEMBERSHIPS

ANNUAL MEMBER

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

PAPER PUBLICATION

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

Process of submission of Research Paper

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

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

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

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

(II) Choose corresponding Journal.

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

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

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

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

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

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

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

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

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

Scope

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

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

2. ETHICAL GUIDELINES

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

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

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

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

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

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

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

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

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

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

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

Please mention proper reference and appropriate acknowledgements wherever expected.

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

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

3. SUBMISSION OF MANUSCRIPTS

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

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

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

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

4. MANUSCRIPT'S CATEGORY

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

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

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

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

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

5. STRUCTURE AND FORMAT OF MANUSCRIPT

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

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

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

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

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

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

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

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

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

(h) Brief Acknowledgements.

(i) References in the proper form.

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

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

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

Format

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

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

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

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

Structure

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

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

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

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

Key Words

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

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

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

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

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

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

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

Acknowledgements: Please make these as concise as possible.

References

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

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

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

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

Tables, Figures and Figure Legends

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

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

Preparation of Electronic Figures for Publication

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

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

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

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

6. AFTER ACCEPTANCE

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

6.1 Proof Corrections

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

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

(Free of charge) from the following website:

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

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

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

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

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

6.3 Author Services

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

6.4 Author Material Archive Policy

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

6.5 Offprint and Extra Copies

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

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

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

TECHNIQUES FOR WRITING A GOOD QUALITY RESEARCH PAPER:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Informal Guidelines of Research Paper Writing

Key points to remember:

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

Final Points:

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

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

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

General style:

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

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

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

In every sections of your document

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

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

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

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

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

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

Title Page:

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

Abstract:

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

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

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

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

Approach:

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

Introduction:

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

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

Approach:

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

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

Procedures (Methods and Materials):

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

Materials:

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

Methods:

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

Approach:

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

What to keep away from

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

Results:

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

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

Content

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

What to stay away from

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

Approach

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

Figures and tables

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

Discussion:

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

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

Approach:

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

Administration Rules Listed Before Submitting Your Research Paper to Global Journals Inc. (US)

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation)
by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

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

Index

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