Global Journal of Computer Science and Technology

discovering thoughts and inventing future

Volume 10 Issue 11 Version 1.0 | Online ISSN: 0975-4172 | Print ISSN: 0975-4350

11 Technology Reforming Ideas

highlights

- Information Service System
- Scalar Multiplication Method
- Key Dependent Cryptosystem
- Data Mining Techniques

October 2010

ENG

Global Journal of Computer Science and Technology

Global Journal of Computer Science and Technology

Volume 10 Issue 11 (Ver. 1.0)

Global Association of Research

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

USA Tax ID: 98-0673427

Global Journals Inc.

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

Sponsors: Global Association of Research
Open Scientific Standards

Publisher's Headquarters office

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

Offset Typesetting

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

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

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

eContacts

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

Pricing (Including by Air Parcel Charges):

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

Editorial Board Members (HON.)

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

From the Chief Author's Desk

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

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

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

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

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

Robust Digital Image Watermarking Scheme Based on DWT and ICA

G.Thirugnanam[1],S.Arulselvi[2]

GJCST Classification
D.4.6, I.5.4, G.1.2

***Abstract*-In recent years, access to multimedia data has become much easier due to rapid growth of the internet. While this is usually considered an improvement of everyday life, it also makes unauthorized copying and distributing of multimedia data much easier, therefore presenting a field of watermarking.Many literatures have reported about Discrete Wavelet Transform (DWT) based watermarking for data security since they are found to be robust against image processing attacks when compared with spatial domain. In this paper, an attempt is made to develop a watermarking scheme based on DWT and extraction using Independent Component Analysis (ICA).FastICA is proposed and implemented for extraction. In this work, proposed DWT with Fast ICA is compared with DWT based Self Reference with Non ICA technique. Simulation results show that proposed technique produces better PSNR and similarity measure. The robustness of the proposed scheme is evaluated against various image-processing attacks.**

Keywords- Authentication, data hiding, digital watermarking, self-reference, wavelet transform.

I. INTRODUCTION

Recently, the tremendous growth of the internet has increased multimedia services, such as electronic commerce, pay-per-view, video–on-demand, electronic newspapers and peer to peer media sharing. As a result, multimedia data can be obtained quickly over high-speed network connections. However, authors, publishers, owners and providers of multimedia data are reluctant to grant the distribution of their documents in a networked environment because the ease of intercepting, copying and redistributing electrical data in their exact original form encourages copyright violation. Therefore, it is crucial for the future development of networked multimedia systems that robust methods are developed to protect the intellectual property right of data owners against unauthorized copying and redistribution of the material made available on the network. Furthermore, it is an important issue to develop a robust watermarking scheme with a better tradeoff between robustness and imperceptibility.

Watermarking techniques [2],[5] can be broadly classified into two categories: Embedding watermarks in the spatial Domain or in the frequency (transform) domain.Many literatures have reported about watermarking based on _ spatial domain with different conventional extractiontechniques [6].It has been found that these techniques produce poor robustness and less PSNR. Transform domain watermarking techniques are more robust, this is due to thefact that when image is inverse wavelet transformed, watermark is distributed irregularly over the image, making the attacker difficult to read or modify. Among the transform domains DWT, based watermarking techniques are gaining more popularity because of superior modeling of Human Visual System (HVS). Recently many literatures have reported that the watermarking schemes based on Discrete Wavelet Transform (DWT) [4],[5]. In Joo et al self-reference image scheme, the extraction process depends on the embedding location *idx,* as well as original image. The above said feature does not provide robust [1] and thus the watermark can be removed easily. Hence, ICA based extraction of watermark found to be better alternate when compared with Non ICA extraction techniques.To overcome the above said problem, an attempt is made in this paper, to develop and implement watermarking scheme based on Discrete Wavelet Transform and extraction of watermark by a blind Independent Component Analysis (ICA) algorithms for digital images. The novelty of the proposed method is that it does not require original image and embedding parameters such as watermark location and strength.Simulation results are presented for various attacks and it is found that proposed method(DWT with Fast ICA) produces high similarity measure and robust to various image processing attacks like jpeg compression, Gaussian noise, cropping, Rotation and Translation.This paper is organized as follows: Section II reviews the Watermark embedding and extraction, Section III discusses the self reference scheme, Section IV presents the Proposed work, In Section V Simulation Results are presented and conclusions are drawn in Section VI.

About[1]- G.Thirugnanam is Senior Lecturer, Dept. of Instrumentation Engineering, Annamalai University, Tamil Nadu, and India. Telephone: 9842385955 email: thirugnanam_me@yahoo.com).
About[2]. Dr.S.Arulselvi is Reader, Dept. of Instrumentation Engineering, Annamalai University,Tamil Nadu ,and India. (Email: ggtt_me@yahoo.com)

II. WATERMARK EMBEDDING AND EXTRACTION

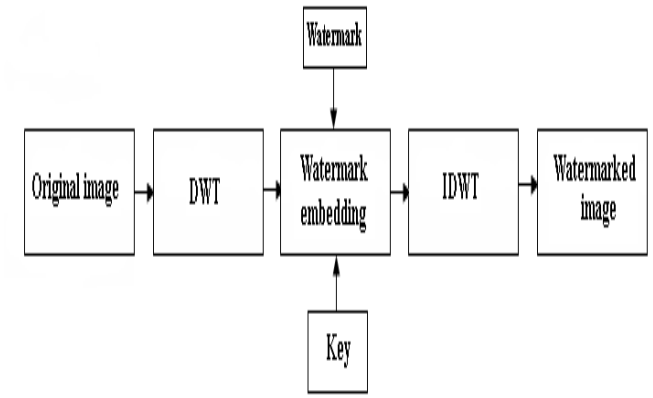


Fig. 1. Watermark embedding

In general, watermark embedding requires an original image and a watermark. In Fig.1 Original image is decomposed to two level using DWT and the watermark is embedded in the HL2 sub-band along with a private key, which is used to hide the watermark. Then IDWT is done to obtain the watermarked image.

1) Decomposition By DWT

Wavelet transform allows the decomposition of the signal in narrow frequency bands while keeping the basis signals space limited. Fig. 2 shows a two level DWT decomposition tree using low pass and high pass analysis filter banks $h(-m)$ and $g(-m)$ respectively [8]. If the level of decomposition is increased, the approximate image will be more stable. However, the complexity increases and the amount of information that can be embedded will be decreased. As a compromised way, the original image is decomposed into two levels. In wavelet analysis, an original image can be decomposed into an approximate image LL1 and three detail images LH1, HL1 and HH1 as shown in Fig. 2. Using wavelet analysis on the approximate image LL1 again, four lower-resolution sub-band images LL2 and three detail images LH2, HL2 and HH2 will be obtained and the approximate image holds the most information of the original image. Others contain some high-frequency information such as the edge details and these detail images can be affected easily by the noise, some common image processing, etc, so they are not stable enough to hide information in them. However, the watermark can be embedded into the approximation coefficient. Thus, the degree of robustness will be improved and integrity of the details which improve the imperceptibility. The main drawback of not embedding the watermark in the LL2 will lead to serious degradation of image quality. Hence, the watermark can be embedded either in HL2 or LH2. In this paper, HL2 sub-band is chosen to embed the watermark.

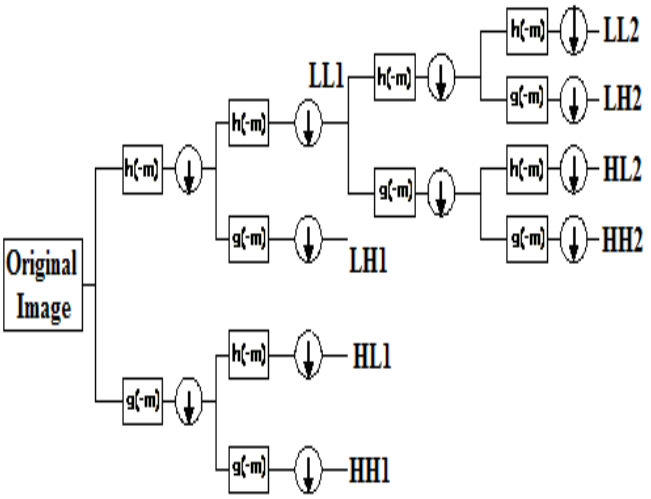


Fig. 2. Two level 2D DWT analysis filter banks

2) Embedding

The watermark embedding procedure is to embed a watermark into the original image in HL2 sub-band obtained from DWT as shown in Fig. 2. The watermark can be perceptible or imperceptible in the watermarked image depending on the applications. For applications, requiring the original image not being distorted, imperceptible watermark is desired. For some other applications, which require displaying the embedded image, a perceptible watermark is preferred. In this paper, imperceptible watermark is obtained.

3) Extraction

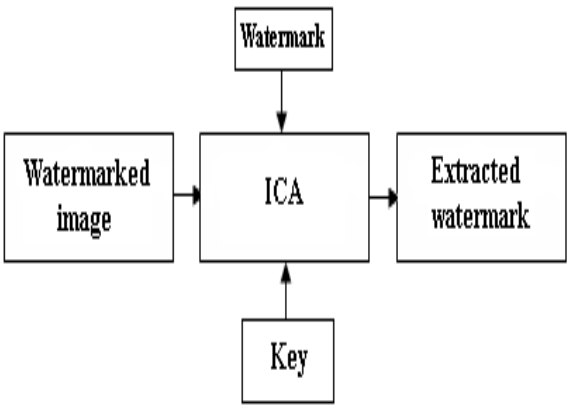


Fig. 3. Watermark extraction

Watermark extraction is used to retrieve the embedded watermark from the watermarked image as shown in Fig. 3. For watermark extraction, a secret key, which is the same used during embedding, is used together with the watermarked image to retrieve the embedded watermark. In this paper, FastICA is proposed to extract the watermark.

4) Independent Component Analysis

ICA is a statistical technique for obtaining independent sources S from their linear mixtures X, when neither the original sources nor the actual mixing A are known. Thisis achieved by exploiting higher order signal statistics and optimization techniques. The result of the separation process is a demixing matrix W, which can be used to obtain the estimated unknown sources, $\overline{S}$ from their mixtures. This process is described by

$$X = AS \rightarrow \overline{S} = WX \quad (1)$$

FastICA algorithm applied in this work for watermark extraction is discussed below:

Aapo Hyvarinen and Erkki Oja have proposed an Fast ICA algorithm and it is based on a fixed-point iteration scheme [7]. The operation of Fast ICA algorithm is outlined as follows:

The mean of the mixed signal X is subtracted so as to make X as a zero mean signal as

$$X = X - E[X] \quad (2)$$

Where $E[X]$ is the mean of the signal?

ii) Then covariance matrix is

$$R = E[XX^T] \quad (3)$$

is obtained and eigenvalue decomposition is performed on it, where E is the orthonormal matrix of eigenvalues of R and D is the diagonal matrix of eigenvalues. Find the whitening matrix, P which transforms the covariance matrix into an identity matrix is given by

$$P = Inv\left(sqrt(D) \times E^T\right) \quad (4)$$

iii) Choose an initial weight vector W, such that the projection $W^T X$ maximizes non-gaussianity as

$$W^+ = E\{X * g(W^T X)\} - E\{g'(W^T)\}W \quad (5)$$

Where g is the derivative of the nonquadratric function. The variance of $W^{+T} X$ must be made unity. Since X is already whitened it is sufficient to constrain the norm of W^+ to be unity.

$$W = \frac{W^+}{\|W^+\|} \quad (6)$$

If W not converges means go back to step (iv).

iv) The demixing matrix is given by

$$W = W^T \times P \quad (7)$$

and independent components are obtained by

$$\overline{S} = W \times X \quad (8)$$

III. Self Reference Scheme

Watermark embedding by Liu et al [1] is shown in Fig. 4. In this scheme, three sub-bands (LH_2,HL_2 and HH_2) are set to zero except LL_2 as stated by Joo et al. After performing inverse wavelet transform, its reference sub-band LL_2' is obtained. The information *idx* of embedding location in the watermark embedding process is obtained by sorting $\left|LL_2 - LL_2'\right|$ .Finally, the watermark information isembedded into the sub-band $LL_2 = LL_2' \pm k \times w(idx(j))$, where j = 1 to 1000, k is a factor for controlling embedding intensity and w is a pseudo-random binary sequence with the length of 1000 bits generated by using a seed, w belongs to [1, -1] and idx is the key. In watermarking extraction process, the original image is required for obtaining the watermark embedding location. According to the embedding location, the watermark can be extracted by comparing the two sub-bands LL_2 and LL_2' . Finally, the extracted watermark is compared with the original watermark by similarity measure. However, the above embedding process is quite time – consuming. Besides, the original image is required in the watermark extraction process, which is impractical in real applications. The two schemes discussed by Joo & Liu, embed the watermark by zeroing of high frequency sub-bands. Hence, the above scheme does not provide robustness as stated by Ting et al [9] and the reason is embedded space where the watermark location can be recovered easily, thus the watermark can be removed or replaced.

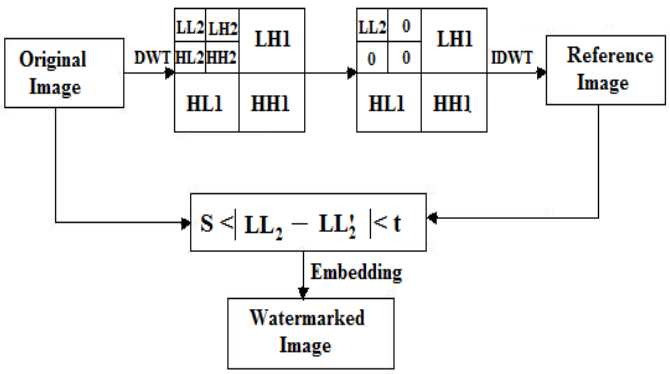


Fig. 4. Watermark embedding process using self reference scheme

IV. Proposed Work

To overcome the above said problems, this paper proposes a watermarking scheme based on DWT and extraction of watermark by blind ICA techniques. This proposed scheme is compared with the existing self-referencenon-ICA extraction technique to evaluate the performance. The novelty of this method is that it does not require original image and embedding parameter such as watermark location and strengthThe watermark embedding process is shown in

Fig. 1. The original image is decomposed into two levels using DWT analysis as shown in Fig. 2. Between the two middle subbands HL_2 is chosen in this work to embed watermark as it provides high PSNR values, also this HL_2 subband provides high robustness and imperceptibility when compared to LH_2. A stochastic model of the cover image is applied to an adaptive watermark by computing NVF with non-stationary Gaussian model [10]. In this case, NVF can be expressed by

$$NVF(i,j) = \frac{1}{1+\sigma_X^2(i,j)} \qquad (9)$$

Where $\sigma_x^2(i,j)$ denotes variance of the cover image in a window centered on the pixel with spatial coordinates (i,j). A masking function named Noise Visibility Function (NVF) is applied to characterize the local image properties, identifying the textured and edge regions where the information can be more strongly embedded. Such high-activity regions are generally highly insensitive to distortion. With the visual mask, the watermark strength can be reasonably increased without visually degrading the image quality.By applying NVF; the watermark in texture and edges becomes stronger than in flat areas. The watermark is embedded using the following equations:

$$I'HL_2(i,j) = HL_2(i,j) + E(HL_2)\alpha\ (1 - NVF(i,j))W(i,j) + \frac{E(HL_2)}{10}.\alpha\ .NVF(i,j)W(i,j) \qquad (10)$$

Where $I'HL_2(i,j)$ are watermarked coefficients, $E(HL_2)$ and $\frac{E(HL_2)}{10}$ denotes the watermark strengths of texture and edge regions, respectively. α is the smoothing factor and E denotes the mean and $W(i,j)$ is the watermark. Then, perform the inverse DWT to retrieve the watermarked image. For watermark extraction, a random key is used together with the watermarked image to retrieve the embedded watermark.

$$X_1 = a_{11}I' + a_{12}W + a_{13}K \qquad (11)$$

$$X_2 = a_{21}I' + a_{22}W + a_{23}K \qquad (12)$$

$$X_3 = a_{31}I' + a_{32}W + a_{33}K \qquad (13)$$

Where a is a mixing matrix, I' is the watermarked image, W is the watermark and K is a random key in the embedding process. Applying the above mentioned ICA algorithms to those mixtures, watermark W is extracted. The watermark is extracted from the watermarked image as shown in Fig. 3.

V. Simulation Results

A gray scale of size 256x256 is considered as original image (Flower image) as shown in Fig. 5. Simulations are carried out using MATLAB software. The original image is decomposed using discrete wavelet transform for two level. The watermark is embedded in the second level middle frequency subband (HL_2) using the embedding equation (10). A binary image of size 64 x 64 is considered as the watermark image (Robut) as shown in Fig. 6. The watermarked image is obtained using two level IDWT. To justify the results, various images are taken and watermark is embedded using DWT to obtain watermarked image and it is shown in Fig. 7(a-e). The Peak Signal to Noise Ratio (PSNR) is calculated between original and watermarked image using the formula (14). PSNR values for various test images are shown in Table 1. From the results, it is observed that the wavelet transform reconstructs the image better when compared to the self reference technique.The quality of the watermarked image is evaluated by calculating the Peak Signal to Noise Ratio (PSNR) betweenoriginal and watermarked image using the formula

$$PSNR = 10\log_{10}\frac{255^2}{MSE}(dB) \qquad (14)$$

Where MSE is the Mean Square Error. The PSNR values calculated for the existing self- reference technique as well as the DWT technique for Flower image is 33.4104 and 41.6662, respectively.

Table 1. PSNR values for various test images

Test Images	Self Reference Technique	Proposed Technique
Flower	33.4104	41.6662
Football	32.8963	39.2112
Peppers	31.1125	36.5659
House	30.4937	34.9548
Moon	35.7625	40.9858

Among the test images, Flower is chosen for simulation. Fig. 8(a-e) show the various image processing attacks like JPEG compression, gaussian noise addition with noise density of 0.5, cropping, Rotation and Translation respectively on Flower image Fig. 9(a–e) shows the extracted watermarks using FastICA from the above mentioned attacks respectively. The similarity measure criteria is calculated between original and extracted watermark using the expression

$$Sim(X, X') = \frac{X.X'}{\sqrt{X'.X'}} \qquad (15)$$

where X is the original watermark and X' is the extracted watermark.

Fig. 5. Original image

Fig. 6. Watermark

(a) Flower

(b) Football

(c) Peppers

(d) House

(e) Moon

Fig. 7(a-e). Watermarked images

(a)Jpeg compression

(b)Gaussian noise

(c) Cropping

(d) Rotation

(e) Translation

Fig. 8(a-e). Various attacks

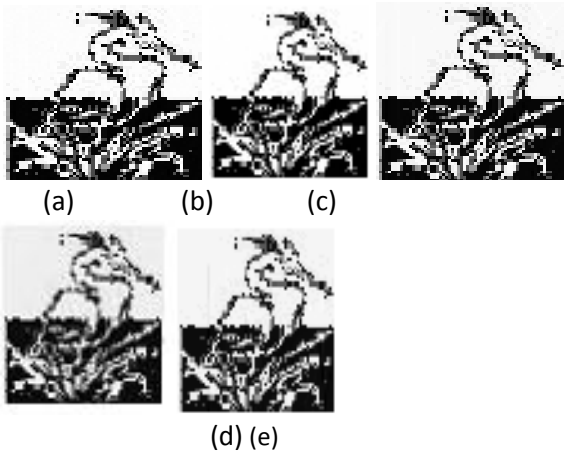


Fig. 9 (a-e). Extracted watermarks

Table 2 compares the performance of PSNR (dB) and Similarity Measure for existing technique (Self Reference with Non ICA) with proposed technique (DWT with Fast ICA) for Flower image. Considerable differences between the two techniques are observed for all type of attacks and it is inferred that DWT with Fast ICA performs better.

Table 2. Performance comparison of existing technique with proposed technique

ATTACKS	PSNR (dB)		SIMILARITY MEASURE	
	Existing Technique	Proposed Technique	Existing Technique	Proposed Technique
JPEG Compression	29.1823	33.3772	0.8916	0.9599
Gaussian Noise	24.8114	25.7510	0.8869	0.9589
Cropping	8.4236	11.4382	0.8824	0.9532
Rotation	13.5321	14.1585	0.8801	0.9538
Translation	22.3428	27.5552	0.8851	0.9571

VI. Conclusion

A Robust Watermarking scheme using DWT with FastICA is presented in this paper and their performance against various attacks are obtained. To evaluate the performance of the proposed technique, it is compared with Self Reference Non ICA technique. The simulation results reveal that the proposed scheme (DWT with Fast ICA) is better interms of PSNR values and similarity measure values.

VII. References

1) Jiang Lung Liu., Der Chyuan Lou., Ming Chang Chang and Hao Kuan Tso. (2006) „A Robust watermarking scheme using self reference image', *computer standards and interfaces,Elsevier,* pp.356-367.
2) Van Schyndel, R.G., Tirkel, A.Z., and Osborne, C.F. (1994) „A Digital Watermark', *International conference on image processing,* pp.86-89.
3) Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I. (1992) „Image coding using wavelet transform', *IEEE Trans. Image Processing,* Vol.1, No. 2, pp. 205-220.
4) Wang, H.J.M., Su , P.C., and Kuo, C.C.J. (1998) „ Wavelet based digital image watermarking', *Opt. Express*, Vol. 3(12), pp.491-496.
5) Kundur, D., and Hatzinakos, D. (1998) „Digital watermarking using multiresolution wavelet decomposition', *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Vol. 5, pp. 2969-2972.
6) Ingemar J.Cox., Matthew L.Miller., and Jeffrey A. Bloom. „ Digital Watemarking and steganography', Second Edition, Morgan Kaufmann Publishers, New York.
7) Hyvarinen, A., Karhunen, J., and Oja, E. „Independent component analysis', First Edition,John Wiley & Sons, Inc., New york.
8) Thirugnanam, G., and Arulselvi, S. (2010) „Wavelet packet based Robust Digital Image Watermarking and extraction using Independent Component Analysis', *International Journal of Signal and Image Processing*, Vol. 1(2), pp. 80-87.
9) Grace C. W. Ting., Bok Min Goi., and Swee Huay Heng. (2008) „Attacks on a robust watermarking scheme based on self reference image', *Computer Standards and Interfaces,* Vol. 30, pp. 32-35.
10) Thai Duy Hien., Zensho Nakao., and Yen-Wei Chen. (2006) „Robust multi-logo watermarking by RDWT and ICA', *Signal Processing*,pp.2981-2991.

Exploring Genetic Algorithm for Shortest Path Optimization in Data Networks

Dr. Rakesh Kumar[1], Mahesh Kumar[2]

GJCST Classification
C.2.1, G.2.2

Abstract **- Internet Service providers (ISPs) are the building blocks for Internet. Due to huge demand of Internet by various business communities and individuals, ISPs are trying to meet the increasing traffic demand with improved utilization of existing resources and application of new technologies. Routing of data packets can affect network resource utilization. A protocol is followed to send packets from source to destination along a path. In intra-domain Internet routing protocol Open Shortest Path First (OSPF) is the most commonly used protocol. As any user can come in and out from the logical topology of network, routing in dynamic network is a challenging one. I have implemented a Genetic algorithm to finds the set of optimal routes to send the traffic from source to destination.**

Keywords- Genetic Algorithm, Chromosome, Crossover, Mutation, Routing

I. INTRODUCTION

Data network routing is a process of transferring packets from source node to destination node with minimum cost (delay-transmission, processing and queuing delay, bandwidth, load, jitter, reliability etc.). Routing is complex in large networks because of the many potential intermediate destinations a packet might traverse before reaching its destination. Hence routing algorithm has to acquire, organize and distribute information about network states. It should generate feasible routes between nodes and send traffic along the selected path and also achieve high performance. Routing in conjunction with congestion control and admission control defines the performance of the network. The weights of links in network are assigned by the network operator. The lower the weight, the greater the chance that traffic will get routed on that link [BLU00]. When one sends or receives data over the Internet, the information is divided into small chunks called packets or datagram's. A header, containing the necessary transmission information, such as the destination Internet Protocol (IP) address, is attached to each packet. The data packets are sent along links between routers on Internet. When a data packet reaches a router, the incoming datagram's are stored in a queue to await processing. The router reads the datagram header, takes the IP destination address and determines the best way to forward this packet for it to reach its final destination. Routing algorithm should havegeneric objective ofrouting strategy to be both dynamically reconfigurable and be based on locally available information. It should also satisfy user quality of service objectives. Some of the methods proposed in achieving these objectives are social evolutionary algorithms, insect metaphors and cognitive packet network. Insect metaphors and cognitive packet network methods use the probabilistic routing table and allow the packets themselves to investigate, report network topology status and performance. Genetic algorithm is one in which the population associated with each node co-evolve to solve the problem.

About[1] : Associate Professor, Department of Computer Science and Applications, Kurukshetra University Kurukshetra, India
(rsagwal@gmail.com, 9896336145)
About[2] : Assistant professor, Department of Computer Science and Engineering, DAV college of engineering & technology, Kanina, Mohindergarh, India
(myadav.kuk@gmail.com, 9813431025)

II. NETWORK ROUTING

Routing is a process of finding paths between nodes in network. Broadly there are mainly two types of routing policies - static and dynamic. In static routing, the routes between the nodes are precomputed based on certain factors for example bandwidth, buffer space etc. and are stored in routing table. All packets between any two nodes follow the same path. When network topology changes the path between two nodes may also change and static routing fails. Hence in dynamic routing policy, the routes are not stored but are generated when required. The new routes are generated based on the factors like traffic, link utilization, bandwidth, jitter, delay etc which is aimed at having maximum performance. For message transmission routing policy may be centralized or distributed. In the case of centralized routing, only centralized node generates routes between any pair of nodes. Centralized routing is not adequate in IP networks as it is required to collect whole network state before route computation, which is very complex task. In distributed routing, each node generates routes independently between pair of nodes as and when required. Other classification of routing policy is optimal routing (global routing) and shortest path routing (local routing) [CAUOO]. Some of the shortest path algorithms are distance vector algorithm and link state algorithm. Each node in the network is of the type store and forward. The link performance may be measured in terms of bandwidth or link delay. The topology of the network may change due to growth in number of nodes, or failure of node. This change in topology should be reflected in the routing table, which in turn helps the routing protocol to generate optimal route for the current state of network. Some of the protocols are Routing Information Protocol (RIP), Interior gateway routing protocol (IGRP), Open source shortest path first (OSPF) and Border gateway protocol (BGP). OSPF is a link state routing protocol used in IP networks which uses shortest path first algorithm to compute low cost route to

destination. Enhanced Interior Gateway Routing Protocol (EIGRP) is an enhanced distance vector routing protocol with optimization to minimize the effect of change in topology and efficient use of bandwidth and processing power at the router. EIGRP uses unequal cost load balancing. Routing process uses routing table at each node to store all the nodes which are at one hop distance from it [CAU00]. It also stores the other nodes (hop count more than one) along with the number of hops to reach that node, followed by the neighbor node through which it can be reached. Router decides which neighbor to choose from routing table to reach specific destination.

III. SHORTEST PATH PROBLEM

The shortest path problem is defined as that of finding a minimum-length (cost) path between a given pair of nodes. The Dijkstra algorithm is considered as the most efficient method for shortest path computation in IP networks. But when the network is very big, then it becomes inefficient since a lot of computations need to be repeated [BIL08]. Also it cannot be implemented in the permitted time [YIN06].

IV. GENETIC ALGORITHM

Genetic Algorithm (GA) is a special kind of stochastic search algorithm that depicts the biological evolution as the problem solving technique. GA works on the search space called population [GOL89]. Each element in the population is called as chromosome. GA begins with randomly selecting set of feasible solution from population. Each chromosome is a solution by itself. Each chromosome is evaluated for fitness and this fitness defines the quality of solution. GA uses adaptive heuristic search technique which finds the set of best solution from the population. New off-springs are generated /evolved from the chromosomes using operators like selection, crossover and mutation. Most fit chromosomes are moved to next generation. The weaker candidates get less chance for moving to next generation. This is because GA is based on the principle of Darwin theory of evolution, which states that the "survival is the best". This process repeats until the chromosomes have best fit solution to the given problem [LAU91]. The summary is that the average fitness of the population increases at each iteration, so by repeating the process for many iterations, better results are discovered. GA has been widely studied and experimented on many fields of engineering. GA provides alternative methods for solving problems which are difficult to solve using traditional methods. GA can be applied for nonlinear programming like traveling salesman problem, minimum spanning tree, scheduling problem and many others.

1) Features of Genetic algorithm

• The most important feature of genetic algorithms is that they are parallel in nature. They explore solution space in multiple directions at once. GA is well suited for solving problems where the solution space is huge and time taken to search exhaustively is very high.

• GA performs well in problems with complex fitness. If the function is discontinuous, noisy, changes over time or has many local optima, then GA gives better results [VIJ08].

• GA has ability to solve problems with no previous knowledge (blind). The performance of GA is based on efficient representation, evaluation of fitness function and other parameters like size of population, rate of crossover and mutation and the strength of selection.

2) Problem definition

The network under consideration is represented as G = (V, E), a connected graph with N nodes. The metric of optimization is cost of path between the nodes. The total cost is the sum of cost of individual hops. The goal is to find the path with minimum total cost between source node V_{src} and destination V_{dest}, where V_{src} and V_{dest} belong to V. This paper presents the efficient on-demand, source initiated routing algorithm using genetic algorithm. Finally data is sent along the generated path.

A. Initialization of routing table

A module is used to generate all possible paths from a given node to all other nodes in the network. Initially, „n' random paths are considered (chromosome). This 'n' defines the population size. These chromosomes act as population of first generation.

B. Optimal paths generation

This module deals with finding the optimal path using genetic algorithm. The input to this module is the set of paths generated. Each path is called as chromosome. As the source node receives „m' (say 10- population size) chromosomes-

(a) Calculate the fitness of each of the chromosome.

The fitness of the chromosome is evaluated as: Fitness = no of hops in path * 10 – total cost of path Number of hops defines the number of intermediate nodes visited along the path from source to destination and total cost is the sum of cost of individual links in the path.

(b) Select best two chromosomes as parents (using some selection method-Roulette wheel)
(c) Perform crossover with probability 0.7
(d) Perform mutation with probability 0.01
(e) Place children in the population and eliminate the worst chromosome having very poor fitness value.
(f) If termination condition is not attained then repeat the steps i. to vi.

Else (Convergence criteria is reached) { • Store the paths for duration t seconds • send data to the destination along the path }

(g) Refresh the path after duration of t seconds to know the current status of dynamic network.

C. Selection

It is a feature of GA for selecting parents for next generation. Current work is based on roulette wheel selection.Roulette wheel selection – In roulette wheel selection, the individual is selected based on the relative fitness with its competitors. This is similar to dividing the

wheel into a number of slices. Fittest chromosomes get larger slice. Some of the other selection methods are rank selection, elitist selection, scaling selection, tournament selection, etc.

D. Crossover

Crossover operator combines sub parts of two parent chromosomes and produces offspring that contains some part of both the parent genetic material. Crossover is mainly of two types namely single point crossover and multipoint crossover. In single point crossover, there is one cross over site and in multipoint crossover there is more than one crossover site. Single point crossover method is simple; it has some problems like formation of cycles when used for routing. Hence it is required to use some of the advanced multipoint crossover techniques to eliminate cycle. Some of the advanced multipoint crossover techniques are Partially Mapped Crossover (PMX), Cycle crossover (CX) and Order crossover (OX) [GOL98, SIV08]. This paper deals with PMX crossover. In Partially Matched Crossover [SIV08], two strings are aligned, and two crossover points are selected uniformly at random along the length of the strings. The two crossover points give a matching selection, which is used to affect across through position-by-position exchange operations.

Consider two strings:

Parent A	4 8 7 |	3 6 5 |	1 10 9 2
Parent B	3 1 4 |	2 7 9 |	10 8 6 5

Two crossover points were selected at random, and PMX proceeds by position wise exchanges. In-between the crossover points the genes get exchanged i.e., the 3 and the 2, the 6 and the 7, the 5 and the 9 exchange places. This is by mapping parent B to parent A. Now mapping parent A to parent B, the 7 and the 6, the 9 and the 5, the 2 and the 3 exchange places. Thus after PMX, the offspring produced as follows:

Child A	4 8 6 |	2 7 9 |	1 10 5 3
Child B	2 1 4 |	3 6 5 |	10 8 7 9

Each offspring contains ordering information partially determined by each of its parents. PMX can be applied to problems with permutation representation. Generated offspring should be validated. Validation is done by checking the offspring with all possible routes. If offspring belongs to all possible routes then its fitness is computed and sent to next operation. If the offspring does not belong to all possible route set, then it is dropped as route does not involve valid connections of nodes in network.

E. Mutation

Crossover operation may produce degenerate population. In order to undo this, mutation operation is performed. Mutation operation can be bit flipping, interchanging, inversion, insertion, reciprocal exchange or others [ALU06]. The paper uses insertion method. In case of insertion a node is inserted at random position in the string. This is because a node along the optimal path may be eliminated through crossover. Using insertion, it can be brought back. Once mutation is completed, the offspring generated by mutation have to be validated with the same process used in crossover.

F. Termination Criteria

It allows the convergence of algorithm. Maximum generations, No change in population fitness and stall generation are considered as algorithm stopping condition. We have taken the maximum number (say 1000) of generations as it will allow algorithm to check, upto what number of generations there is improvement in chromosome fitness. A second stopping criterion is until some chromosome reaches a specified fitness level. As the optimal solution is generated using GA, data is transmitted along that path. There may be change in topology of network as some nodes may join the network or some nodes may leave the network or some nodes may fail. Under these circumstances the optimal path may no more be the shortest. Hence the network has to be refreshed at every t seconds and new routes may be generated.

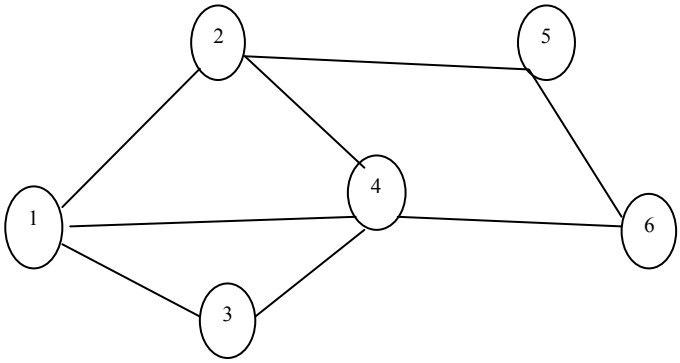


Figure 1 : Sample Network Topology

The Cost on links is given in table 1:

	1	2	3	4	5	6
1	999	**5**	**3**	**7**	999	999
2	**5**	999	999	**3**	**5**	999
3	**3**	999	999	**3**	999	999
4	**7**	**3**	**3**	999	999	999
5	999	**5**	999	999	999	**3**
6	999	999	999	**2**	**3**	999

Table 1: Cost on linksThe values 999 represents that there is no direct link between these nodes. 999 is a big value as compared to other costs. During implementation only small values are considered for path computation.

V. Results

Current work is based on network consisting 6 nodes. Initially 15 random chromosomes are generated, out of which best ten are considered for 1st generation. At each generation the chromosomes are validated and best fit chromosomes are sent to next generation. It is found that fitness value improves at each generation for chromosomes.

Generate 15 random chromosomes

Chromosome	Delay	Fitness
1 4 2 5 6 3	18	32
1 2 5 6 4 3	18	32
1 4 6 5 3 2	12	18
1 2 4 3 5 6	11	19
1 2 4 3 6 5	11	19
1 2 5 4 3 6	10	10
1 4 2 3 5 6	10	10
1 4 6 3 5 2	9	11
1 2 4 5 3 6	8	12
1 2 3 4 5 6	5	5
1 2 3 6 4 5	5	5
1 2 6 4 5 3	5	5
1 2 6 4 3 5	5	5
1 3 5 6 4 2	3	7
1 6 3 4 5 2	0	0

Generation 1

Chromosome	Delay	Fitness	Nodes visited
1 4 2 5 6 3	18	32	5
1 2 5 6 4 3	18	32	5
1 4 6 5 3 2	12	18	4
1 2 4 3 5 6	11	19	4
1 2 4 3 6 5	11	19	4
1 2 5 4 3 6	10	10	3
1 4 2 3 5 6	10	10	3
1 4 6 3 5 2	9	11	3
1 2 4 5 3 6	8	12	3
1 2 3 4 5 6	5	5	2

Generation 2

Chromosome	Delay	Fitness	Nodes visited
1 3 4 2 5 6	17	33	5
1 3 4 6 5 2	16	34	5
1 4 6 5 3 2	12	18	4
1 2 4 3 5 6	11	19	4
1 2 4 3 6 5	11	19	4
1 2 5 4 3 6	10	10	3
1 4 2 3 5 6	10	10	3
1 4 6 3 5 2	9	11	3
1 2 4 5 3 6	8	12	3
1 2 3 4 5 6	5	5	2

We have taken population size of 10 in first generation. By selecting the chromosomes based on roulette selection and application of GA operators generations are performed. After the path to all nodes from source node 1 is computed, the set of paths to a specific node will be displayed. Let the destination node is node 6. Following is the set of paths from node 1 to node 6. The optimal path returned is 1, 3, 4, and 6 with delay factor of 8.

Source	Destination	Delay	Route
1	6	13	1 2 5 6
1	6	10	1 2 4 6
1	6	9	1 4 6
1	6	18	1 4 2 5 6
1	6	8	1 3 4 6

Table2: Routes to the destination 6

VI. Conclusion

Genetic algorithm is used for routing in packet switched data networks in current work. They explore solution space in multiple directions at once. GA is well suited for solving problems where the solution space is huge and time taken to search exhaustively is very high. As the size of network increases, the possible solutions for transferring data between two nodes increase. Adding of few new nodes in the network increases the size of search space exponentially. So, GA is well suited for routing problem as it explores solution space in multiple directions at once and less chances to attain local optimum.GA has ability to solve problems with no previous knowledge. The performance of GA is based on efficient representation, evaluation of fitness function, population size, crossover rate, mutation probability and the selection method. The proposed algorithm works on initial population created by some other module, access fitness, generate new population using genetic operators and converges after meeting to specified termination condition. Current work can be improved by using some intelligent approach for populating routing table and using better crossover, mutation probabilities and enhancing it to support for load balancing.

VII. Refrences

1) (ALU06) Aluizio F. R. Araújo et. al. (2006): Multicast Routing Using Genetic Algorithm Seen as a Permutation Problem, Proceedings of the 20th International Conference on Advanced Information Networking and Applications (AINA'06)
2) [BIL08) Bilal Gonen (2008): Genetic Algorithm Finding the Shortest Path in Networks, 2008
3) [BLU00) Black U. (2000): IP Routing Protocols, RIP, OSPF, BGP, PNNI & Cisco routing protocols, Prentice Hall
4) (CAU00) Cauvery, N. K. & Viswanatha, K. V. (2000): Routing in Dynamic Network using Ants and Genetic Algorithm, International Journal of Computer Science and Network Security, VOL.9 No.3, March 2000
5) (GOL89) Goldberg E. David (1989): Genetic Algorithms in search, optimization and machine learning, Pearson Education
6) (LAU91) Laurence Davis (1991): Hand book of Genetic Algorithms, 1991
7) [SIV08) Sivanandam S.N. & Deepa S.N. (2008): Introduction to Genetic Algorithms, 2008

8) (VIJ08) Vijayalakshmi K & Radhakrishnan S (2008): Dynamic Routing to Multiple Destinations in IP Networking using Hybrid Genetic Algorithm (DRHGA), International Journal of Information Technology, Vol 4, No 1, PP 45-52, 2008
9) (YIN06) Yinzhen Li, Ruichun He, & Yaohuang Guo. (2006): Faster Genetic Algorithm for Network Paths, The Sixth International Symposiumon Operations Research and Its Applications (ISORA'06), pp.380–389

Enhancement of Exon Regions Recognition in Gene Sequences Using a Radix -4 Multi-valued Logic with DSP Approach

D. Venkat Reddy, D. V. Paradesi Rao And E.G.Rajan

GJCST Classification
J.3

***Abstract*-Numerous levels of concepts perform logical designand logical representations in an efficient manner. In typical and quantum theories of computation, Binary logic and Boolean algebra occupies an imperative place. But they havethe limitation of representing signals or sequences by using either binary '1' or '0'. This has major drawbacks that the neutralities or any intermediate values are ignored which are essential in most of the applications. Because of the occurrence of such situations it is the need of the hour to look into other alternative logics in order to fulfill the necessities of the user in their respective applications. The binary logic can be replaced by Multi-Valued Logic (MVL), which grabs the positions of the major applications because of the ability to provide representation by using more than two values. As most of the significant applications are based on the logical sequences, the multi-valued logic shines because of its thriving feature. Genomic signal processing, a novel research area in bioinformatics, is one of the foremost applications which involve the operations of logical sequences. It is concerned with the digital signal representations and analysis of genomic data.Determination of the coding region in DNA sequence is one of the genomic operations. This leads to the identification of the characteristics of the gene which in turn finds out an individual's behavior. In order to extract the coding regions on the basis of logical sequences a number of techniques have been proposed by researchers. But most of the works utilized binary logic, which lead to the problem of losing some of the coding regions and incorrectly recognizing non-coding regions as the coding regions. Hereby, we are proposing an approach for recognizing the exon regions from a gene sequence based on the multi-valued logic. In this approach, we have utilized fourlevel logical system, termed as quaternary logic for the representation of gene sequences and so that we recognize theexon regions from the DNA sequence.**

Keywords-Multi-valued logic (MVL), Quaternary logic, spectral component, gene sequence, exon regions, exon regions recognition

I. INTRODUCTION

Sundry number systems have been developed rightfrom the evolution of computers. These systems concentrate on simplifying the basic mathematical operations so as to assist in making the computer more powerful [6]. Complex mathematical operations in computing systems are carried out in terms of logical operations that make the computation easier. For any signal xin the circuit, we indicate its logic values before and after a clock transition as $x(t)$ and $x(t0)$ respectively [3]. Logic design is carried out in an effective way at a variety of levels of abstraction. In general it starts with a "word" description of the design problem, which is later represented as a truth table from which a logic function is formulated. The majority of the logic design has been on the basis of the binary logic because of its intuitive relationship to the binary states of electronic switches, ON and OFF or abstractly 0 and 1. The similar mind set has been carried over to logic design by means of bio-molecules in spite of them being free from this restriction [2]. On the basis of the number of states of the technology (components) used [2] the type of logic used will depend [2]. Digital circuit design has traditionally been linked with binary logic where the two logic levels are represented by two discrete values of current, voltage or charge [5]. Due to the intuitive relationship to the binary states of electronic switches, ON and OFF or abstractly 0 and 1[2] a large amount of the logic design are based on the binary logic.Binary logic and Boolean algebra plays a vital role in typical and quantum theories of computation. But, in Binary logic there is a possibility for only two outputs which denotes either a true condition (1) or a false condition (0). When a neutral or an intermediate value has to be represented as an output in the application then it is referred to a partially correct value either 1 or 0. It is that binary will submit a value or in any way fade and die i.e., the neutralities are ignored in this logic. These types of situations have raised the alarm to search for other alternative logics in order to satisfy the user requirement in their respective applications. The MVL is considered to be a potential substitute to Binary logic [9]. For the last couple of decades, multiple-valued logic (MVL) has attracted significant attention, particularly among circuit and system designers. MVL circuits allow more than two levels of logic and depending on the number of allowed levels, we may have ternary (base = 3) or quaternary (base =4) logic styles. MVL seek out to enhance the information processing efficiency of a circuit by transmitting more information on each signal line than simple binary logic and by implementing complex functions of the inputs in a single gate [4]. MVL circuits can minimize the number of operations necessary to implement a specific mathematical function and further, have an advantage in terms of reduced area [5], which in turn reduces parasitics linked with routing

About-Associate Professor Department of E.C.E
venkatreddyphd@gmail.com

and offers a higher speed of operation [1]. Compared with the conventional binary logic, the information density in MVL is much higher [1]. Some of the recent research works related to the MVL is given as follows. Yi Jin et al. [11] have proposed a theory known to be the Decrease-Radix Design. And on the basis of their theory, the regulations of building multi-valued logic operation units were offered. The theory laid down a solid foundation forthe design of re-constructible logic units in ternary optical computers in addition to any other multi-valued computers. The key of the theory was that if the physical states to represent information incorporated a special state "D", then any of the $n(n \times n)$ n-valued LUs could be realized by means of the combination of the $n \times n \times (n-1)$ operationbasic-units (OBU) according to the DRD theory, where "D" is a special physical state named by the authors. It would result in A while the state D operated with any state A. M.H. Nodine et al. [12] have provided an overview of methods proposed in order to implement multi-valued logic in CMOS and then portrayed Intrinsity's patented Fast reg Technology as a mature methodology for silicon implementation of multi-valued logic. Fast Technology was on the basis of three fundamental characteristics comprising the utilization of (1) footed NMOS transistor domino logic, (2) multi-phased overlapping clocks, and (3) 1-of-N encoding of MVL signals. In order to offer additional opportunities for power optimization, the concepts of null value as well as mutex properties were introduced, presenting additional challenges for MVL representation as well as synthesis. A high-level design method of multiple-valued arithmetic circuits is proposed by Y. Watanabe et al. [13]. Their method utilized a cell-based approach with a dedicated hardware description language called ARITH. By means of ARITH, they could portray and verify any binary/multiplevalued arithmetic circuits in a formal manner. The ARITH description may perhaps be transformed into a technologydependent netlist in binary/multiple-valued fused logic. The process of transforming the netlist into a physical layout pattern was automatically carried out by an off-theshelfplace-and-route tool. They presented a particular cell library comprising a multiple-valued signed-digit adder and its related circuits with a 0.35mum CMOS technology, and illustrated that their method possibly will synthesize a 32times 32-bit parallel multiplier in multiple-valued currentmode logic from an ARITH description. M.H.A. Khan et al. [14] have offered a heuristic algorithm for concurrent variable ordering as well as quaternary Galois field expansion selection in order to build optimal quaternary Galois field decision diagram (QGFDD). They also give you an idea about way of flattening the QGFDD in order to generate QGFSOP expression. They have written Java program for the purpose of building QGFDD for multioutput quaternary functions and offered experimental results. Navi et al. [9] have offered two versions of current mode 1-bit adder. The related propagation delays of them are 70ps, 150ps respectively. Those adders made use of only 6 transistors. The chip density advantage of the multi-valued approach was significant. They carried out simulations through the HSPICE in a 0.18µm technology at 27 centigrade; with 1.8 volt supply voltage. The simulation results demonstrated that they attained a remarkable improvement in terms of transistor count, chip area as well as propagation delay. The minimum number of transistors reported in current mode was 11 and it is 10 in voltage mode. That is they accomplished 40% performance in terms of transistor count and the improvement in speed is about 2.5%. B.J. Falkowski et al. [10] have offered classification of novel fastest quaternary linearly independent transforms. They were defined recursively and also had consistent formulas linking their forward and inverse transform matrices. Their transform matrices' properties and calculation example were revealed. The computational costs of the calculation for offered transforms as well were discussed. The experimental results were made known and compared with the well known quaternary arithmetic transform.The architecture of Chinese abacus adder is offered by Shu-Chung Yi et al. [15]. As high radix of adder might minimize the number of carry propagation, their Chinese abacus adder might achieve high-speed operation. The simulation results of their works were compared with CLA (Carry Lookahead) adder. The delay of the 8-bit abacus adders are 22%, 17%, and 14% less than those of CLA adders for 0.35µm, 0.25µm, and 0.18µm technologies, respectively. The power consumption of the abacus adders were 30%,34%, and 60% less than those of CLA adders for 0.35µm, 0.25µm, and 0.18µm technologies, respectively. The utilization of Chinese abacus approach resulted in a competitive technique with regard to conventional fast adder. Luis E. Cordova et al. [16] have presented an approach in order to generate molecular electronics systems by means of introducing a multi-valued programmable logic (MVPL) block for the purpose of modeling the characteristics lately observed experimentally in few molecular structures so as to achieve circuit robustness on a molecular substrate. They demonstrated that given the experimentally observed characteristics of few types of molecular electronics substrates, they possibly will adopt a multi-valued logic system for the logic blocks, by this means partially avoiding the necessity to incorporate specialized fault detection or fault tolerant circuitry at the molecular level that would be essential in binary logic circuits. They evaluated their MVPL model against the other work presented in the literature with respect to the efficiency of the formulation in order to understand computing architectures in the face of high defects in the self-assembled substrate. MVL has a number of advantages when compared to the existing binary systems. Improving binary logic levels to ternary, penta and quad levels, more advanced processing values are obtained in a variety of computing applications. Genomic signal processing is one of the most noteworthy application areas where MVL can be implemented to overcome drawbacks faced while computed with binary logic. A major challenge for genomic research is to establish a relationship among sequences structures and functions of genes. DNA sequence

is an emblematic string of letters 'A', 'C', 'G' and 'T'. The segments of DNA molecule, called genes are recognized to carry valuable information in their protein coding regions (exons) and are responsible for protein synthesis. In eukaryotes, exon regions are segmented by non-coding regions (introns), while in prokaryotes these regions are continuous [7]. The pivotal problem of gene identification in eukaryotes is distinguishing exons, from introns and intergenic regions [8]. Enormous selections of techniques are employed in separating the sequence from the DNA, until now binary logic is being utilized more normally to extract thissequence. In the mean time, while computing with binary logic the sequence obtained may contain noises impregnated with them which make it difficult to distinguish between the exons and introns. Even after the noise being removed, the coding regions are only recognized with a stress. Also when the magnitude of exon is improved in order to enhance the strength of the signal it will alternately increase the magnitude of the intron which also causes a drawback in the identification of the exons from the introns. Hereby, we are proposing an approach which overcomes the mentioned drawbacks in identifying the protein coding regions based on binary logic. The proposed approach utilizes quaternary logic, radix-4 multi-valued logic which replaces the binary logic. Utilizing the four-level logic in the conventional technique of coding region identification, we obtain the exons more clearly. In order to replace the binary by quaternary values, we are proposing a simple, but more effectual logical conversion technique which converts the binary represented gene sequence into quaternary indicators. Hence, the intention of extracting the information region from any of the gene sequence can be achieved by our proposed quaternary level based coding region identification. The rest of the paper is organized as follows: Section 2 gives a brief introduction about the multi-valued logic and the quaternary logic which is utilized in our approach and Section 3 briefs some fundamental ideas about the genomic signal processing. Section 4 succinct the conventional binary logic based DSPmethod for exon regions identification and Section 5 details the proposed quaternary logic based DSP method with sufficient formulations and illustrations. Section 6 discusses about the implementation results and section 7 concludes the paper.

II. MULTI-VALUED LOGIC (MVL)

Traditional calculi are merely two valued for any proposition. MVL are logical calculi wherein there are above two truth values [22]. Numerous forward-thinking efforts dedicated to the MVL synthesis have been made in recent years, in specific, however an effective methodology for MVL design is until now an open challenge [19]. Being able to minimize the number of interconnection lines or nets and enhance their information content, MVL turn out to be quite attractive [17]. In multi-valued logic, the connectives as well as the rules for building formula are those utilized in classical logic, and the disjunction, conjunction and negation of formulas are defined by max, min operations in addition to the complementation to 1, respectively [18]. There are numerous advantages of such logic systems as well as circuits when compared with the binary ones. The major advantages of MV digital and computer systems are: enhanced speed of arithmetic operations realization, superior density of memorized information, improved usage of transmission paths, minimizing of interconnections complexity and interconnections area, lessening of pin number of integrated circuits and printed boards, possibilities for easier testing [20]. One can attain a more cost-effective way of exploiting interconnections by means of a larger set of signals over the same area in MVL devices, permitting easy implementation of circuits. In MVL devices, the noise advantage of binary logic is preserved. The higher radix in use is the ternary (radix-3) as well as the quaternary (radix-4) [21].

1) Quaternary logic

Quaternary logic is very appropriate for encoded realization of binary logic functions by grouping 2-bits together into quaternary digits. This sort of quaternary encoded reversible realization of binary logic function necessitates half times input/output lines than the original binary reversible realization. As the number of input/output lines is minimized, this quaternary encoded realization of binary logic function causes the circuit more compact and manageable, in particular for the quantum technology, where the cost of qudit (quantum digit) realization and qubit (quantum bit) realization are almost similar [23]. Quaternary identity logic circuit works as a buffer and is utilized in designing quaternary D flip-flop. Quaternary identity logic circuit is composed of thermometer code circuit, EXOR gate, 2 bias inverters, Ntype transmission gate, and P-type transmission gate [24]. Quaternary logic is of more interest when compared to other types of multi-valued logic because of the simplicity of signal grouping by two bits [19].Quaternary Signed Digit numbers are represented by means of 3-bit 2's complement notation. Every number can be represented by

$$X = \sum_{i}^{n} x_i 4^i$$

where xi can be any value from the set {3, 2,1,0,1,2,3} in order to produce an apt decimal representation [25]. The quaternary logic makes use of 0, 1, 2 and 3 logic levels. Figure 1 illustrates quaternary logic levels

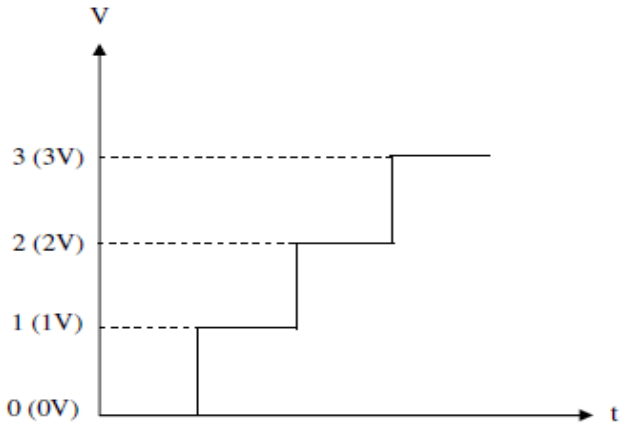


Fig. 1: 3V supply level for quaternary logic representationWhen the binary logic is used to represent any number, only 2*i*combinations are possible (when *i* number of bits is used for representation). But in Quaternary logic, the bit level representations are up to *i* 4 combinations. While applying the value of *i* = 3, the total no of combinations that can be represented is 64. This makes the quaternary logic more strong and advantageous when applied for signal or sequence representations. Taking all the features into concern, we have employed the logic for gene sequence representation in the application of identification of protein coding regions.

III. Genomic Signal Processing

With the vast amount of genomic and proteomic data that is available in the public domain, it is becoming more and more significant to be able to process this information in ways that are helpful to humankind. In this context, raditional as well as modern signal processing methods have played a significant role in these fields [26]. Gene identification is one of the most vital tasks in the study of genomes [31]. The engineering discipline that studies the processing of genomic signals is broadly classified as Genomic signal processing (GSP) [30]. The definition of the Genomic Signal Processing can be given as the analysis, processing and the utilization of genomic signals in order to obtain biological knowledge as well as with that knowledge system-based application are devised. Due to the most important role played in genomics by transcriptional signaling and the related pathway modeling, it is only natural that the theory of signal processing be supposed to be utilized in both structural and functional understanding.

The aim of GSP is to combine the theory and methods of signal processing with the global understanding of functional genomics, with special importance on genomic regulation [27]. DNA is the building block of all life on this planet, from single cell microscopic bacteria to more advanced creatures like humans [29]. DNA topology is of primary importance for an extensive range of biological processes. Because of the topological state of genomic DNA is of importance for its replication, recombination and transcription, there is an immediate interest to acquire information regarding the supercoiled state from sequence periodicities. Identification of dominant periodicities in DNA sequence will help understand the significant role of coherent structures in genome sequence organization [30]. Genome sequencing is figuring out the order of DNA nucleotides, or bases, in a genome that frame an organism's DNA. Genome sequences are big in size and are capable of ranging from several million base pairs in prokaryotes to billions of base pairs in eukaryotes [29]. The DNA is made up of long sequences of four kinds of nitrogen containing bases {A, C, G, T}. These sequences are grouped in coding regions – exons (of the eukaryotic genes) and non-coding regions – introns (a variety of regulatory regions such as promoters, enhancers, silencers, long repeats with apparently no function, etc.). The coding regions are translated into proteins, whereas the immense majority of the non-coding regions appear to have no biological function whatsoever [28]. Accurate prediction of exon regions is a research problem at present being addressed [7]. So as to be successful, a gene finding algorithm has to incorporate good indices for the protein coding regions [31]. Additionally processing and analyzing this information are of major importance. The volume of genomic data is expanding at an enormous as well as still growing rate, while its basic properties and relationships are not so far fully understood and are subject to continuous revision. This data is stored, managed, and analyzed on a huge diversity of computing systems, from small personal computers which makes uses many disk files to supercomputers operating on large commercial databases [32].

IV. Conventional DSP Approach for EXO Regions Recognition

As we have talked about previously, there are four nucleotides (or bases) comprised in the strands of DNA. They are designated by the characters A, T, C,and G. A haracter string composes of these four bases. And such a character string can be mapped to four signals. The conventional DSP methods for coding region identification utilize the binary signals to represent the signals. To be brief, the string which is composed of four bases is mapped into four binary signals. The value of „l' is taken by the signal bA(n) in the case if A is present in the DNA sequence at index n . But if it is not the case that is if A is absent at index n the value of „0' is taken. For instance, bA(n) for the DNA segment „CGTCGTGGAA' is given as 0000000011. In the same manner the signals bT (n) ,bG (n) and bC (n) can be acquired. After that the DFT of bA(n) , BA(f) over W samples is found. . In the same manner it is possible to obtain the DFT of bT (n) ,bG (n) and bC (n) , termed as BT (f) , BG (f) and BC (f) respectively. Period-three behavior is noticed in several genes and it is also found that is very much helpful in recognizing the coding regions [33]. In addition, several researchers have observed that the period-3 property to be a good (preliminary) indicator of gene location [26]. For this reason, the (f = N / 3) –DFT coefficient magnitude is

frequently considerably larger when compared to the surrounding DFT coefficient magnitudes. And this corresponds to a coding region inside the gene. Based upon the gene [33] this effect differs and be
able to be fairly pronounced or fairly weak. A figure that can be utilized in order to measure the total spectral content S(f) of a DNA character string at frequency f is defined as the sum of the magnitude of the DFT values of the four binary nucleotide sequences. Observe that a calculation of the DT at the single point f = N / 3 is adequate. The window can after that be slid by one or more bases and S(N / 3) recalculated. Therefore, we obtain a picture of how S(N / 3) evolves along the length of the DNA sequence. It is essential that the window length W be adequately large (typical window sizes are a few hundreds, e.g., 351, to a few thousands). On the other hand a long window implies longer computation time, and in addition compromises the base-domain resolution in predicting the exon location [26]. On the other hand the non-coding regions in the DNA spectrum at3 2p are not wholly suppressed by the conventional DSP.

Therefore, a non-coding region may be mistakenly recognized as a coding region [34]. To overcome these shortcomings, we have replaced the binary logic by MVL for mapping the sequences in the proposed approach and so that exact coding regions are identified effectively.

V. PROPOSED QUATERNARY LOGIC BASED EFFECTIVE DSP APPROACH

In this paper, MVL based processing of genomic signals in order to recognize the exon regions are proposed. Here, MVL used is four-level logic i.e. quaternary logic for identifying the coding regions in a gene sequence. The steps involved in the quaternary logic based coding region identification are depicted in the figure 2. The approach replaces the binary logic by quaternary logic by means of a conversion technique. Thus the obtained quaternary indicators are subjected for the further process of exon identification in the gene sequence.

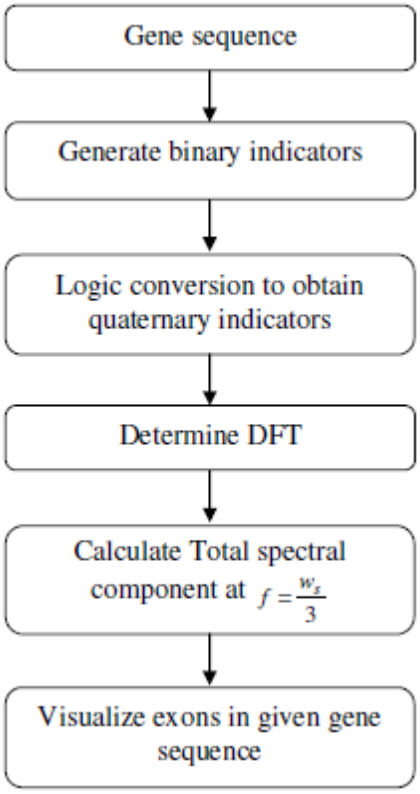


Fig 2: Proposed Quaternary logic based DSP approach for exon regions recognition

As an initial process which is depicted in the figure 2, we are utilizing the binary logic representation of the gene sequences. Let the sequence be gs , which is the combination of nucleotide bases A (Adenine), G (Guanine), C (Cytosine) and T (Thiamine) having the length of l . The binary indication of the any such sequences is given as

$$b_T(n) = \begin{cases} 1; & if \quad g_s(n) = 'T' \\ 0; & else \end{cases} \tag{1.b}$$

$$b_A(n) = \begin{cases} 1; & if \quad g_s(n) = 'A' \\ 0; & else \end{cases} \tag{1.a}$$

$$b_G(n) = \begin{cases} 1; & if \quad g_s(n) = 'G' \\ 0; & else \end{cases} \tag{1.c}$$

$$b_C(n) = \begin{cases} 1; & if \quad g_s(n) = 'C' \\ 0; & else \end{cases} \tag{1.d}$$

In equation (1) ,b_A (n), b_T (n), b_G (n) and b_C (n) are the binary indications of the DNA sequence g_s representing the nucleotide bases 'A', 'T', 'G' respectively where j= 1,2,3…l Then the binary indicators are applied to a logical conversion which converts

the binary indicators into quaternary indicators. Logic conversion: The conversion technique, heart of the proposed approach presented here converts the binary indicators into quaternary indicators. The conversion procedure takes a few steps which are given as follows.
Initially the indices of the binary indicators are obtained as
$[A_P] << \Delta b_A(n)$; if $b_A(n) = 1$ (2.a)
$[A'_P] << \Delta b_A(n)$; if $b_A(n) = 0$ (2.b)In equation (2), A_P and A'_P are the index vectors of $b_A(n)$ which have the index values of logical '1's and '0's respectively. Then the four different logical levels are assigned to each of the binary values constituted by the indicators as

$$A_V = \begin{cases} V_1 ; & if \ \ A_P(n) \% 2 = 0 \\ V_3 ; & else \end{cases} \quad (3.a)$$

$$A'_V = \begin{cases} V_0 ; & if \ \ A'_P(n) \% 2 = 0 \\ V_2 ; & else \end{cases} \quad (3.b)$$

Equation (3) represents the conversion of the binary indicator values to the quaternary logical values.V_0, V_1, V_2 And V_3 are the four logical levels utilized in our approach which directly represents the values 0,1,2 and 3 respectively.$A_V(x)$ and $A'_V(x)$ are the vectors which have he quaternary values of logical '1' and logical '0' of the binary indicator respectively. As per the index taken from the binary indicator, the values should be concatenated in order to get the final quaternary indicators of the DNA sequences. This can be performed as

$$A_1(n) = \begin{cases} A_V(x); & if \ \ A_P(x) = n \\ 0 \quad ; & else \end{cases} \quad (4.a)$$

$$A_2(n) = \begin{cases} A'_V(x); & if \ \ A'_P(x) = n \\ 0 \quad ; & else \end{cases} \quad (4.b)$$

In equation (4), if $A_P(x) = n$ and $A'_P(x)$ n are not satisfied, then the value of x is not incremented. It remains until the condition gets satisfied. Then by performing the addition operation between the vectors A1 and A2 , the final quaternary indicator for the nucleotide base 'A' is obtained. This can be given as
$qA(n) = A1 + A2$ (5)
where, qA(n) is the quaternary indicator for the nucleotide base which is converted from the corresponding binary indicator . In similar fashion, the quaternary indicators for other nucleotide bases qT (n) , qG (n) and qC (n) are obtained. Henceforth, we use the quaternary indicators for the purpose of the identifying the protein coding region. After obtaining the four quaternary indicators, the spectral content is calculated. The spectral content calculation and the further process of coding region identification are performed for a window of sequences, which is sliding in nature. Let, the window performs a single sliding movement having the size of ws . Then the total number of windows used in our approach is given by

$n_W = l$ --- $W_S + 1$ (6)

where, nw is the total number of windows as we have chosen single sliding movement and ws as window size. For each further process, the sequence covering by the window is applied. For spectral component calculation, we have to determine the DFT for $q_A^{(z)}(m)$ the quaternary indicator sequence at z^{th} window which can be given as

$$Q_A^{(z)}(f) = \sum_{m=0}^{w_s-1} q_A^{(z)}(m) \exp\left(-j\frac{2\pi f}{w_s} m\right);$$
$$0 \le f \le w_s - 1 \quad (7)$$

Here, $q_A^{(z)}(m)$ represents the quaternary indicator sequence covered by the window z of size W_S , $1 \le z \le n_w$.
Hence the DFT of the quaternary indicator, $Q_A^{(z)}(f)$ representing the base 'A' is determined for all the windowed sequences which are sliding. Likely, the DFT will be determined for all the other quaternary indicators, $q_T^{(z)}(n)$ AND $q_G^{(z)}(n)$ so that we obtain $q_C^{(z)}(n)$ $Q_G^{(z)}(f)$And $Q_C^{(z)}(f)$ Then the total spectral content of the sequence at certain frequency f is given as

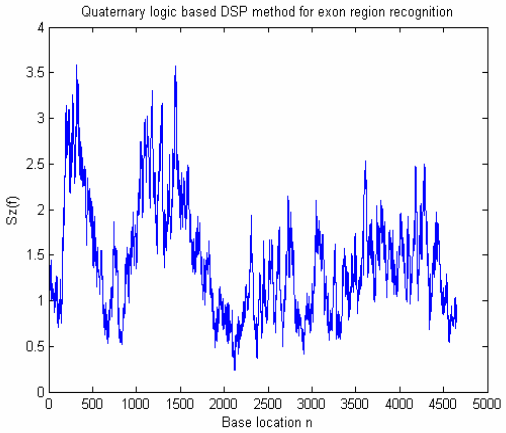
$$S_z(f) = |Q_A^{(z)}(f)|^2 + |Q_T^{(z)}(f)|^2 + |Q_G^{(z)}(f)|^2 + |Q_C^{(z)}(f)|^2 \quad (8)$$

The spectral content, hereby, calculated using the equation (8) considers all the nucleotide bases of DNA sequences. Then, with the aid of the period-3 behavior, the exons are identified from the spectral content. This is due to the fact that the distribution of bases in the exons, which are the integer multiple of 3 exhibit the period-3 behavior. The reason of the period-3 behavior is because of the presence of 3-nucleotide code structure in protein coding region. The period-3 behavior of coding protein region refers to the maximum of Fourier power spectrum (FPS) at the position Of 1/3 fequency [35]. Therefore, it can be decided that the Value of $S_Z(f)$ is maximum at $f = w_s / 3$ while a coding region is there. As a result, after obtaining the total spectral content for all the gene sequences based on the window size, the period-3 behavior is taken. The peaking magnitudes appearing in the total spectral components $S_Z(f)$ give the exon regions. Hence , by the approach we can easily identify most of the exon regions occupying in the gene sequences mainly because of the thriving contribution of the quaternary logic. As the usage of quaternary logic makes the exon regions more dominating in the spectrum rather than the usage of binary logic, the exon

regions are not affected by any of the disturbances. However, the proposed approach struggles in identifying the exon which does not exhibits period-3 behavior. Though, the approach has this drawback, it is very strong in identifying the exact exon regions instead of mistakenly taking the introns because of the effective work of quaternary logic.

VI. Results and Discussion

We have implemented the proposed quaternary logic based coding region recognition in the working platform of MATLAB (version 7.8) from which we have visualized the performance of the approach. For performance visualization, we have utilized DNA sequences of two different organisms, namely, Brucella Suis and Caenorhabditis elegans (C. elegans). As discussed earlier, we have replaced
the binary logic by quaternary logic in the conventional DSP
method of coding region recognition from the mentioned gene sequences. The size of the sliding window, we have chosen for taking the sample sequence is 351 (i.e.ws =351). Then, as per the approach procedure, the spectral components are determined for each window of sequences. As the gene sequences are very huge, it takes too much time to recognize all the coding regions throughout the sequences. Hence we have taken two different samples of sequences and so we have recognized the exon regions within the given sequences. The spectral component for the two different samples of sequences, 5000 and 20,000 for the gene Brucella suis is shown in the figure 3(a) and figure 3(b)
respectively.



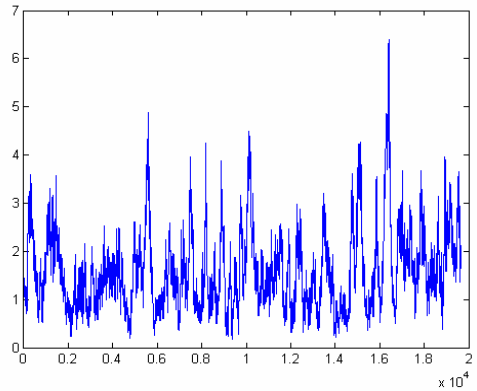


Fig 3: Normalized spectral content of the gene sequence Brucella suis for samples (a) 5000 and (b) 20,000

From the figure 3, the exon regions of the gene Brucella suis
has been visualized clearly. As discussed earlier, the exon regions are nothing but the peaking magnitudes of the spectral content of the gene sequences. It is clear that the proposed approach increases the magnitude of the exon regions rather than the intron regions and so we have obtained all the exon regions which exhibits period-3 behavior. Thus the proposed approach overcomes the major drawback of incapability of recognizing the exons in binary logic based coding region recognition. We have tested the approach using the chromosome III of C. elegans also for seeking different results. The spectral results obtained for C. elegans chromosome III is illustrated in the figure 4.

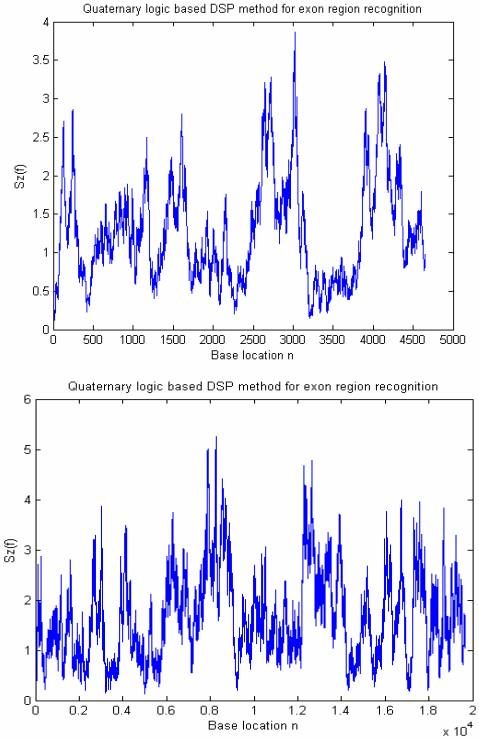


Fig 4: Normalized spectral content of the gene sequence C. Elegans chromo some III for samples (a) 5000 and (b)20,000

In similar fashion, we have obtained the spectral component regions of the C. elegans gene sequence which is constituted by the exon as well as the intron regions. The regions at which the magnitude of the sequence peaks have been recognized as the exon regions exhibiting period-3 behavior.

Figure 4 (a) gives the exon regions of the sequence for a sample of 5000 and figure 4 (b) gives the exon regions for a sample of 20,000. A comparison is made between both the binary and the proposed quaternary based DSP method for exon regions recognition for the two gene sequences, Brucella suis and C. elegans. The comparison plot of spectral component distribution between the conventional and the proposed approach has been given in the figure 5.

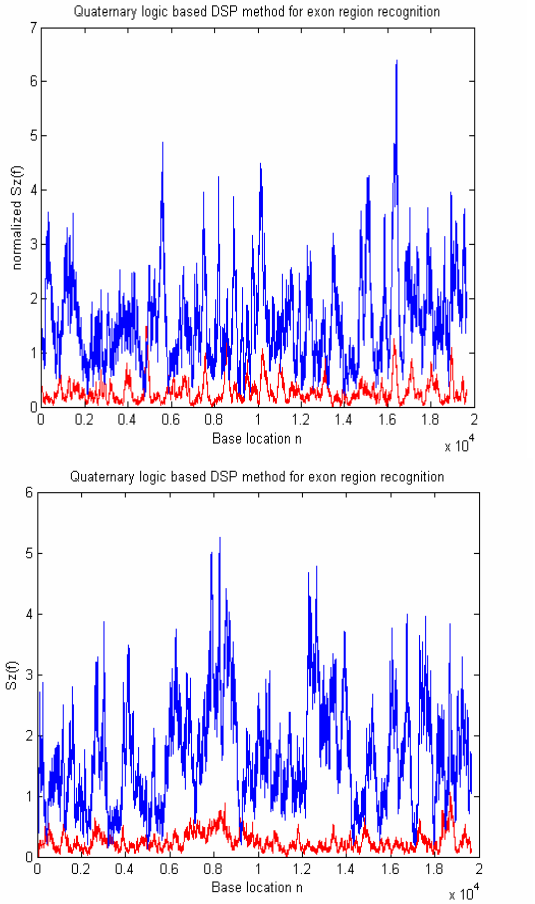


Fig 5: A comparison between the spectral components obtained from the conventional binary logic based DSP method and quaternary logic based DSP method for coding region recognition in the gene (i) Brucella suis and (ii) C. Elegans chromosome III .

The comparison provided between the conventional binary based DSP method and the proposed quaternary based DSP method clearly illustrates the performance difference in recognizing the exon regions. Conventional method shows only a very small spike in the exon regions, but the proposed approach makes a huge peak in the same. This makes clear that the proposed approach is more effective in recognizing the exon regions mainly because of the performance of the quaternary logic.

VII. Conclusion

Being an alternative to binary logic, in this paper, we have proposed a radix-4 MVL based approach for exact identification of exons from the gene sequence. With the aid of the fundamental DSP technique, we have developed the quaternary logic based DSP approach for exon regions recognition in gene sequences. The proposed approach just replaces the binary logic by quaternary logic in the conventional DSP method to identify the coding region in DNA sequence. Because of the utilization of quaternary logic in DSP method of coding region recognition, the magnitude of the coding regions has been increased heavily. This makes the identification of coding region from the gene
sequences more comfortable. It has been well known that the coding region exhibits period-3 behavior and so it peaks when the gene sequences are applied for spectral content calculation. The approach has performed more effectively that it identified the exact exon regions and restricted the introns from domination. This makes the incorrect decision of taking the introns as exons have been mitigated. Hence, the proposed quaternary logic based DSP approach for recognizing the coding region in DNA sequence is more effective rather than binary logic based approach and so we can identify the exact coding regions and not no-coding regions.

VIII. References

1) Wu Gang, Cai Li, and Li Qin, "Ternary logic circuitdesign based on single electron transistors", Journal of Semiconductors, Vol 30, Issue 2, February 2009.
2) H.A. Aleem and F. Mavituna, "Multiple Valued Logic as a Design Tool for Synthetic Biology",University of Manchester, UK.
3) Xunweiwu, Bangyuan Chen and Massoud Pedram, "Power Estimation in Binary CMOS Circuits Based on Multiple-valued Logic", Multi. Val. Logic., Vol
00, pp. 1-17, 2000.
4) H. L. Chan, S. Mohan, Pinaki Mazumder, and George I. Haddad, "Compact Multiple-valued Multiplexers Using Negative Differential Resistance Devices",IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 31, NO. 8, AUGUST 1996.
5) Arijit Raychowdhury and Kaushik Roy, "A NovelMultiple-Valued Logic Design Using Ballistic Carbon Nanotube FETs", Proceedings of 34th International Symposium on Multiple-Valued Logic, pp: 14-19, 2004.
6) Mehdi Hosseinzadeh, Somayyeh Jafarali Jassbi, and Keivan Navi, "A Novel Multiple Valued Logic OHRNS Modulo rn Adder Circuit", International Journal of Electronics, Circuits and Systems, 2007.
7) Mahmood Akhtar, "Comparison of Gene and Exon

Prediction Techniques for Detection of Short CodingRegions", International Journal of Information Technology, Vol. 11, Issue 8, pp: 26-35, 2005.

8) Achuthsankar S. Nair and Sivarama PillaiSreenadhan, "A coding measure scheme employingelectron-ion interaction pseudopotential (EIIP)",Bioinformation, Vol 1 Issue 6, pp: 197–202, 2006.
9) Navi, Foroutan, Mazloomnejad, Bahrololoumi,Hashemipour and Haghparast, "A Six Transistors Full Adder", World Applied Sciences Journal, Vol.4,No.1, pp.142-149, 2008.
10) B.J. Falkowski, Cheng Fu, "Classification of Fastest Quaternary Linearly Independent Arithmetic Transforms," Proceedings of 38th International Symposium on Multiple Valued Logic, pp. 169-173,22-24 May, Dallas, TX, 2008.
11) Yi Jin, Jun-Yong Yan and Kai-zhong Zuo, "Decrease-radix design principle for carrying/borrowing free multi-valued and application in ternary optical computer," Science in China SeriesF: Information Sciences, Vol.51, No.10, pp. 1415-1426, October 2008.
12) M.H. Nodine and C.M. Files, "A MatureMethodology for Implementing Multi-Valued Logicin Silicon," Proceedings of 38th InternationalSymposium on Valued Logic, pp. 2 - 7, 22-24 May,2008.
13) Y. Watanabe, N. Homma, K. Degawa, T. Aoki and T.Higuchi, "High-Level Design of Multiple-ValuedArithmetic Circuits Based on Arithmetic DescriptionLanguage," Proceedings of 38th InternationalSymposium on Multiple Valued Logic, pp.112-117,22-24 May, Dallas, TX, 2008.
14) M.H.A. Khan, N.K. Siddika and M.A. Perkowski,"Minimization of Quaternary Galois Field Sum ofProducts Expression for Multi-Output QuaternaryLogic Function Using Quaternary Galois FieldDecision Diagram," Proceedings of 38th InternationalSymposium on Multiple Valued Logic, pp. 125-130,22-24 May, Dallas, TX, 2008.
15) Shu-Chung Yi, Kun-Tse Lee, Jin-Jia Chen, Chien-Hung Lin and Chuen-Ching Wang," The ne architecture of radix-4 Chinese abacus adder", I proceedings of IEEE International symposium onMultivalue logic, pp. 12, 17-20 May, 2006.
16) Luis E. Cordova, and James P. Davis, "TowardsAchieving Molecular Circuit Robustness Using a Multi-Valued Programmable Logic Model,"Proceedings of IEEE International Workshop on Design and Test of Defect-Tolerant NanoscaleArchitectures, May 1-2, Palm Springs, CA, USA,2005.
17) Chung-Yu WU and Hong-Yi Huang, "Design andApplication of Pipelined Dynamic CMOS TernaryLogic and Simple Ternary Differential Logic", IEEEJournal of Solid-State Circuits, Vol. 28, Issue. 8, pp:895-906, 1993.
18) Xiang Li and Baoding Liu, "Hybrid Logic andUncertain Logic", Journal of Uncertain Systems Vol.3, Issue. 2, pp: 83-94, 2009.
19) Ashur Rafiev, Julian P. Murphy, Danil Sokolov andAlex Yakovlev, "Conversion Driven Design ofBinary to Mixed Radix Circuits", IEEE InternationalConference on Computer Design, pp: 410-415, 2008.
20) Dusanka Bundalo, Zlatko Bundalo, Aleksandar Iliskovic and Branimir Djordjevic, "Architecture andDesign of Multiple Valued Digital and Computer Systems", 1'st Balkan Conference in Informatics,2003.
21) S.L. Hurst, "Multiple-valued logic – its status and itsfuture," IEEE Transactions on Computers, vol. C-33,December 1984, pp. 1160-1179.
22) D. Venkat Reddy, Ch.D.V. Paradesi Rao, E. G.Rajan, "Sequential Circuits in the Framework O (2n+1)-ary Discrete Logic", IJCSNS International Journal of Computer Science and Network Security,Vol. 8 Issue.7, July 2008.
23) Mozammel H. A. Khan, "Reversible Realization of Quaternary Decoder, Multiplexer, and DemultiplexerCircuits", in proc. of 38th IEEE Intl. Symposium on Multiple Valued Logic, pp: 208-213, 22- 24 May 2008, Doi: 10.1109/ISMVL.2008.33.
24) Sangmi Shim, Gisoo Na, Seungwoo Park andSeunghong Hong, "Design of Q-IDEN D Flip-FlopUsing RS-latch", in proc. of International Journal onComputer Science and Network Security, Vol. 6,Issue. 9A, September 2006.
25) Abdul Ahad S.Awwal, Syed M. Munir, A.T.MShafiqul Khalid, Howard E. Michel and Oscar N. Garcia, "Multivalued Optic Parallel Computation Using an Optical Programmable Logic Array", Journal of Informatica, Vol 24, pp: 467-473, 2000.
26) P. P. Vaidyanathan and Byung-Jun Yoon, "The roleof signal-processing concepts in genomics andproteomics", in proc. of Journal on Franklin Institute,vol. 341, no. 1-2, pp: 111- 135, 2004.
27) Edward R. Dougherty, Ilya Shmulevich and MichaelL. Bittner, "Genomic Signal Processing: The SalientIssues", EURASIP Journal on Applied SignalProcessing, vol. 2004, pp: 146- 153, Jan. 2004.
28) Ioan Oprea, Sergiu Pasca and Vlad Gavrila, "Metho of DNA Analysis Using the Estimation of theAlgorithmic Complexity", in proc. of LeonardoElectronic Journal of Practices and

Technologies,no.5, pp: 53-66, July-December 2004.

29) Sara Nasser, Adrienne Breland, Frederick C. HarrisJr., Monica Nicolescu, and Gregory L. Vert, "FuzzyGenome Sequence Assembly for Single and Environmental Genomes", in proceeding of Fuzzy Systems in Bioinformatics and Computational Biology , vol. 242, Springer, pp: 19-44, 2009.
30) A.M. Selvam, "Quantumlike Chaos in the FrequencyDistributions of Bases A, C, G, T in Human Chromosome1 DNA", Apeiron, vol. 11, no. 3, July2004.
31) Jianbo Gao, Yan Qi, Yinhe Cao, and Wen-wen Tung,"Protein Coding Sequence Identification by Simultaneously Characterizing the Periodic and Random Features of DNA Sequences", Journal of Biomedicine and Biotechnology, Vol 2, pp: 139–146,2005.
32) Edward R. Dougherty, Ilya Shmulevich, Jie Chen, and Z. Jane Wang, "Genomic Signal Processing andStatistics", EURASIP Book Series on Signal Processing and Communications, Vol 2, 2005.
33) S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," Comput. Appl. Biosci., vol. 13, no. 3, pp.263–270, 1997. http://bioinformatics.oxfordjournals.org/cgi/reprint/13/3/263
34) Trevor W. Fox, Alex Carreira, "A digital signal processing method for gene prediction with improvednoise suppression", EURASIP Journal on Applied Signal Processing, Volume 2004, Pages: 108 – 114, 2004, ISSN:1110-8657, DOI 10.1155/S1110865704309285.
35) Yuan Xin TIAN, Chao CHEN, Xiao Yong ZOU, JianDing QIU, Pei Xiang CAI1, Jin Yuan MO,"FourierPower Spectrum Analysis of Exons for the Period-3Behavior", Chinese Chemical Letters,Vol.16,No.7,pp939942,2005.http://www.imm.ac.cn/journal/ccl.ht

An Efficient Word Matching Algorithm for off Line Text

Sattyam Kishor Mishr[1], Manish Pande[2]

GJCST Classification
F.2.2 , E.2, K.8.1

***Abstract*-Word processing application that are used today perform a wide variety of jobs; the most challenging of them is to search for a given sequence of characters of a word is called string matching. To find all the appearances of the pattern P in the text T, word matching problem is used. Pattern P neither encloses white space nor anticipates and pursue by space. It is predicated that our text is offline. For solving the word matching problem, word searching algorithm (WSA) has been propounded. WSA works by splitting the offline text in to number of tables and search the pattern by using the brute force manner. The main drawback of WSA algorithm is, to search each appearances of the pattern by the brute force manner in each table. Due to this WSA increase the number of comparisons of the word. This paper proposed an algorithm, that is developed to reduce the number of comparisons to search the pattern in the offline text.**

Keywords- Algorithm; string matching; hashing; offline searching; word searching

I. Introduction

In all string matching problems, all the appearances of the pattern P in the text T are to be reported. Many algorithms have been proposed for solving the word matching problem. To find all the appearances of the pattern P in the text T, word matching problem is used. Pattern P neither encloses white space nor anticipated and pursue by space. Ibrahiem et al in 2008[1] has propounded an algorithm called word searching algorithm (WSA) for solving the word matching problem by splitting the offline text in to number of tables in the pre-processing phase. In the searching phase, they perform character to character comparisons with of the same length in the table by brute force manner.In existing work one more algorithm come for solving the word matching problem called modified word searching algorithm (MWSA) [2]. In this algorithm they use the efficient hash function called SDBM hash function [3, 4] for finding the hash value of each word in the text and the hash value of the pattern P. In WSA has main problem with the searching phase, because the searching phase encloses the brute force manner [1]. In brute force manner they perform character to character comparisons with the same insearching phase. We use SDBM hash function due to very length. In MWSA, they use the SDBM hash function which reduces the time taken to search the word. SDBM hash function has very less chance of collision even in a very large text. [2, 3, 4] In this paper, the modification in WSA algorithm is done by using the SDBM hash function in pre- processing phase and balanced binary search tre less chance of collision even in a very large text. Proposed algorithm has two phases that works as follows: In pre-processing phase: whole text T is read and split into nearly equal sizes. For every part we create a table. Table enclosesstarting position of each word in first column and hash value (computed by the SDBM hash function) of every word in second column for each part of the text. For each table balanced binary search tree is created, which include the hash value of every words of each part in its nodes. The pre-processing phase is done only once. In searching phase: the hash value H of pattern P is computed using the same hash function. This hash value H is searched in the balanced binary search tree. If a match is found for a hash value, then the appearances of the pattern P is verified The main drawback of WSA algorithm is that after splitting the text into number of tables it searches the pattern in the row corresponding to pattern length by brute force manner [1, 2]. Drawback of WSA algorithm is eliminated by proposed algorithm by using the SDBM function [2] and balanced binary search tree [5, 6]. In proposed algorithm, only one table in pre-processing phase and use balanced binary search tree in searching phase. Our algorithm has taken less number of comparisons than the WSA algorithm to search the pattern in the offline text.This paper is organized as follows: In section 2, it is explained existing word matching algorithms. In section 3 it is explained our propose algorithm in section 4, simulation and comparative results are explained. Finally we conclude in section 5.

About[1]. Computer Science and Engineering Department Maulana Azad National Institute of TechnologyBhopal, India satyam_satyam@live.com
About[2]. Computer Science and Engineering Department Maulana Azad National Institute of TechnologyBhopal, India manishpandey@manit.ac.in

II. Existing Word Matching Algorithms

1) *Word Searching Algorithm (WSA)*

In this section we explain the recent optimal word searching algorithms (WSA) and modified word searching algorithm (MWSA) [1, 2]. In WSA algorithm, the given text is predicated to be offline. This algorithm has two phases that works as follows: The first phase which is pre-processing phase starts with reading and splitting the text T into k number of equal parts depending on size of the text T and constructing k number of tables with two columns for each part of the text. The first column contains length of the word and the second contain the starting position of each word in the text. The start positions of the words will be in the same row for the same length. Once the table is constructed, it is sorted in ascending order using the length of the words as a key for sorting. This phase is done once. The second phase is searching for a specific pattern. The algorithm calculates the length of the pattern and search for the same length in

the tables starting from the first table. If the length does not exist in the first table then the algorithm searches for the word in the next table and so on. If the current table is the last table, then a message will indicate that the pattern does not exist. On the other side if the length finds in the table, then the algorithm will report the words in the text using the stored start position in the table and begin to compare. If a full match occurred then occurrence of the pattern is reported. But if word does not find the same character, then the algorithm will move to the next start position and compare again. [1]

2) Modified Word Searching Algorithm (MWSA)

The main drawback of WSA algorithm was that it searches each appearances of the pattern by the brute force manner in each table. To improve WSA algorithm, MWSA algorithm was proposed. In MWSA, the whole text T is splits in to k equal parts according to size of the text T. For each part two tables A and B are created. Tables A and B have two column in each. In pre-processing phase: each part of the text is read. While reading, lengths of the words are computed and according to length of the words, starting positions of the words are stored in the table A and each starting position in table A is mapped to a corresponding hash value in table B. Starting position and hash value of all words of same length are stored in same row corresponding to length of the word in its table. Table B is sorted row wise using insertion sort with hash value H as the key to sort and perform corresponding change in table A, as hash value is mapped to a unique starting position in table. In the searching phase: When a pattern P is read its hash value H is computed using same hash function and its length is computed. This hash value H is searched in each tables in rows that corresponds to words of same length. Binary search is used to search the hash value in table B. When the hash value of pattern matches with the hash value of a word in table B, then using starting position of word in table A, the word is matched character to character with the pattern P. If a complete match occurs, the occurrence of pattern P in text T is reported. [2, 3, 4]

III. Proposed Algorithm

In this section, we explain the proposed algorithm for modification of algorithm explained in section 2. In our proposed algorithm: it is predicated that text is offline. Proposed algorithm works in two phases as follows.

1) Preprocessing Phase

The whole text splits into k equal parts according to size of the text T. For each part, one table having two columns is created. In the pre-processing phase: each part of the text is read. While reading starting position, hash value of words are computed and stored in the table. The SDBM hash function based on bit shifting is used to compute the hash value H of the words in text T and pattern P. In SDBM hash function the hash value H is computed as follows:

$$H = (H<<6) + (H<<16) - H + ch$$

Where ch is the ASCII value of each characters in the word w, H is initialized as zero, "<<" is a bitwise left shift operator. The SDBM is a standard hash function which has very less chances of collision, even in a very large text. The SDBM implementation is based on an algorithm by P.A. Larson known as "Dynamic Hashing" [4]. Algorithm 1 shows the preprocessing phase.

2) Searching Phase

In pre-processing phase we construct a table and store the starting position and hash value in the table. Table is stored row wise. In a split part if same hash value is found then starting positions are stored in same row for same hash value. In searching phase, balanced binary search tree is constructed for the hash value stored in table. When a pattern P is read then its hash value H is computed using SDBM hash function. While computing the hash value of pattern P, same hash function is used which used to compute the hash value of words stored in the table. Hash value H of the pattern P is searched in the balanced binary search tree. When the hash value H of the pattern matches with the hash value of a word in tree, then using starting position of word in table, the word is matched character to character with the pattern P. If complete match occurs, the occurrence of pattern P in text T is reported. A pattern may exist any number of times in the text so we solve this problem while reading the text. We store the starting position of words of same hash value in the same row of the table. For each hash value of same word matches with hash value of the pattern P, character to character comparison is done and occurrence is reported, if complete match occurs. If hash value of the pattern is not found then it moves to the next table and same process is performed. Algorithm 2 shows the searching phase.The drawback of WSA algorithm given by Ibrahiem et al [1] was that when a pattern P is to be searched, they compute the pattern length and search the table in row corresponds to same length in a brute force approach (i.e. every word in a row in a table is compared character to character which is very inefficient way of searching). In proposed algorithm, it is optimized the searching by tree searching the hash value instead of character comparison [2]. WSA is optimized in proposed algorithm by using a single table in pre-processing phase and balanced binary search tree in searching phase. Our proposed algorithm has less complexity than previous algorithms and requires very less number of character comparisons than the previous one. Figure.1 shows the flow chart of the proposed algorithm.

3) SDBM Hash Function

In proposed algorithm for word searching problem, we use key distribution for every words in the text. For key distribution we use SDBM hash function. We use SDBM hash function in our proposed algorithm because it has very less chance of collision, even in a very large text.SDBM hash function is based on bit shifting. In SDBM hash function the hash value H is computed as follows:

$$H = (H<<6) + (H<<16) - H + ch$$

Where ch is the ASCII value of each characters in the word w, H is initialized as zero, "<<" is a bitwise left shift operator.The SDBM hash function has a good overall distribution for many different data sets. It works well in

situations where there is a high variance in the MSBs of the elements in a data set. It was found to do well in scrambling bits, causing better distribution of the keys and fewer splits. It also happens to be a good general hashing function with good distribution. [2]The SDBM implementation is based on an algorithm by P.A. Larson known as "Dynamic Hashing" [4].

Algorithm 1 Pre-processing phase: Table creation using starting position and hash value

Input: Offline text T

1. BEGIN
2. Split _text (K_i, T)
3. FOR i=0 to n
4. Creat_table (K_i [x][y])
5. FOR x=0 to n
6. FOR y=0 to n
7. INSERT (K_i[x][y]=starting position(w_i),hashvalue(w_i))
8. END FOR
9. Creat_Binary_search_tree(hash value (w_i))
10. END FOR

Algorithm 2 Searching phase: search pattern through hash value

Input: pattern (P)

1 compute_hash_value (pattern(P))
2 search (hash value (P))
3 IF (hash value (P)=hash value(w_i)) THEN
4 GOTO creat_table (K_i [x][y])
5 FIND starting position(w_i)
6 MATCH (pattern(P) with Word(w_i))
7 ELSEIF
8 GOTO (Creat_Binary_search_tree(hash value(w_i)))
9 HALT

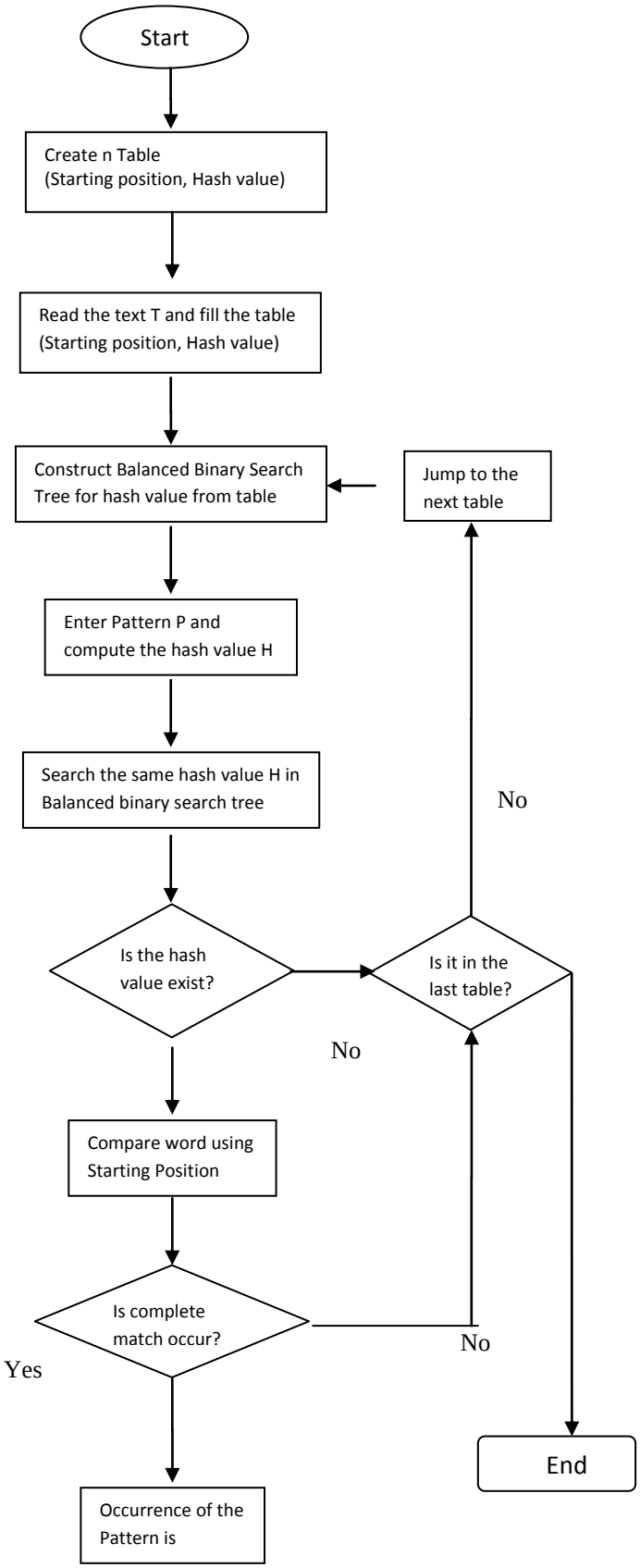


Figure.1 Flow chart of proposed algorithm

IV. Simulation and Results

We have implemented both algorithms WSA and proposed algorithm in C, compiled with GCC 4.2.4 compilers on CORE^{tm} 2 duo 2.66 GHz machine with 4 GB RAM, running Open suse 11.0 Comparison between word matching algorithm. The pattern and text are chosen from ASCII character set randomly.

1) Simulation

For the comparison, we compare WSA algorithm with propose algorithm. Now, we will make a comparison between the proposed algorithm one of the most famous algorithm in such area which is the WSA algorithm to find out the improvement in the number of character comparisons that is done in each algorithm. As an example, we'll take the next paragraph to apply the algorithm with the patterns; "sensor", "processing". Wireless sensor network have emerged as an important application of
1 10 17 25 30 38 41 44 54 66
The ad hoc networks paradigm like monitoring physical environment and
69 73 76 80 89 98 103 114 123 135
These sensor networks have limitations of system resources like battery
137 143 150 159 164 176 179 186 196 201
Power communication range and processing capability Lowprocessing power and
209 215 229 235 239 250 261 265 276 282
In first phase of the simulation each line of this paragraph forming one part of the text, so we have four parts and each part have one table. Each table has two columns. First column has hash value of each word, which is computed by the SDBM hash function and second column has starting position of each word. This is called Preprocessing Phase of the algorithm. The tables will be as following:

Table1. 1st part of the text

Hash Value	Starting Position
3249753982	1
4180602746	10
1691730926	17
385796840	25
2805811025	30
6363218	38
6363213	41
3634954210	44
220492368	54
7281591	66

Table2. 2nd part of the text

Hash Value	Starting Position
700788817	69
6363203	73
866478108	76
2492020741	80
2065330219	89
3573516087	98
1595209896	103
2646868407	114
187585459	123
808581975	135

Table3. 3rd part of the text

Hash Value	Starting Position
1652765827	137
4180602746	143
2492020741	150
385796840	159
1325831833	164
7281591	176
817658959	179
506364869	186
3573516087	196
1906312109	201

Table4. 4th part of the text

Hash Value	Starting Position
2706407781	209
4119183190	215
384379389	229
808581975	235, 282
2081762867	239, 265
1028192824	250
635155988	261
517759365	276

In second phase of the simulation, a pattern "sensor" is taken to search in the text. The hash value of the pattern H is computed by using the SDBM hash function. Now H make searched with the each hash value in each table by using balanced binary search tree. At the time of searching the every comparison is counted. Balanced binary search tree is generated for each table one by one. As the hash value is found the pointer go to the starting position to match the pattern. If the pattern is matched the match report is displayed otherwise negative report is displayed and pointer move to the next comparison in the tree.

Case I. If we have same hash value of the word in the same table then in preprocessing phase the starting position of each hash value will be saved in the same row as shown in Table4. In Table4, Hash value 808581975 has two different starting positions in the same table but saved at the same row.

Case II. If any case pattern does not match with the word at the same starting position then the negative report is displayed and the pointer will be moved to next starting position in the same row.

Case III. Number of comparisons are counted as the pointer move to the next to next node of the balanced binary search tree for searching the hash value of the pattern as well as the number of pattern match at the different starting positions for the same hash value.

2) Comprative Results

As a comparative illustration between the proposed algorithm and WSA algorithm, the character comparison is taken as a parameter. We show the output results as show in the next figure2 and figure3 for the pattern "sensor" and "processing". The results show that how much comparison is done for given patterns. 20 and 40 maximum characters have been taken for the first and second illustration respectively.For the pattern "sensor" which has hash value 4180602746, only 2 comparisons have been taken in proposed algorithm for the first search and 7 comparisons have been done in WSA algorithm for the first search within 20 maximum characters. After taken 40 maximum characters, 10 comparisons and 26 comparisons have been done in propose algorithm and WSA algorithm respectively.For the pattern "processing" which has hash value 2081762867, is not find with in 20 maximum characters comparisons , but after taken 40 maximum characters, 16 and 17 character comparisons have been done for first and full pattern search in proposed algorithm and 41 and 42 comparisons have been done for first and full pattern search in WSA algorithm.

Table5.Comparative results after number of character comparison

	Algorithm comparison based on character comparison			
	Pattern	Algorithms	Number of comparisons	
			First pattern comparison	Full patterns comparison
1	"sensor"	Word Searching Algorithm (WSA)	7	26
		Proposed Algorithm	2	10
2	"processing"	Word Searching Algorithm (WSA)	41	42
		Proposed Algorithm	16	17

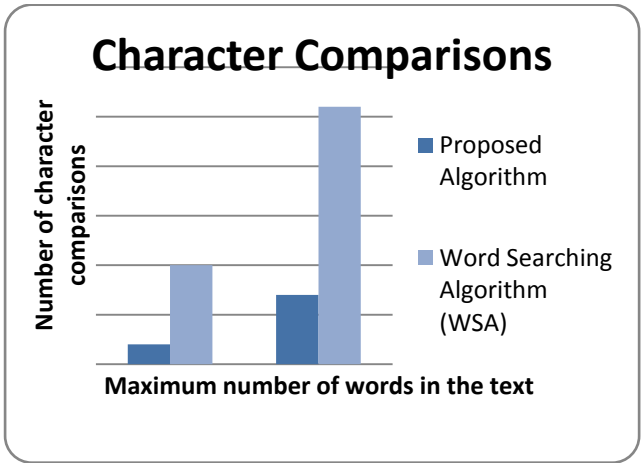


Figure2. Character comparisons in view point of number of character comparisons (Y axis) against maximum number of words in the text
(X axis) for the pattern "sensor".

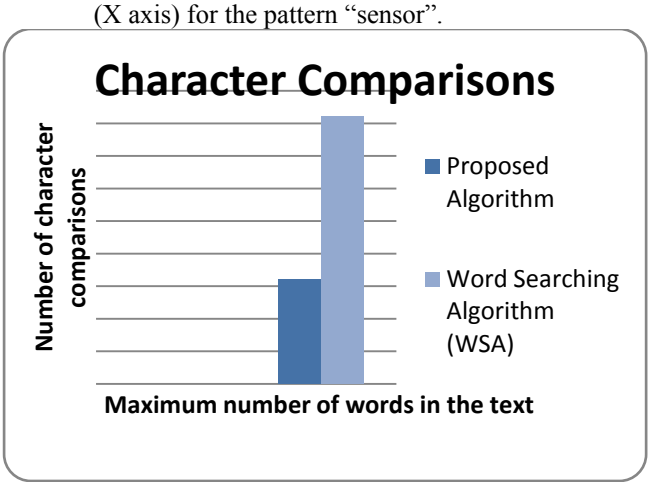


Figure2. Character comparisons in view point of number of charactercomparisons (Y axis) against maximum number of words in the text (X axis) for the pattern "processing".

V. Conclusions

We introduce a new algorithm, which reduces the comparison of word searching algorithm (WSA). From above result it is clear that the proposed algorithm has taken less comparison to find the pattern than word searching algorithms (WSA). In future work proposed algorithm can be compared with the modified word searching algorithm (MWSA) and other pattern matching algorithms.

VI. References

1) Ibrahiem M. M. Emary and Mohammed S. M. Jaber, "A New Approach for Solving String Matching Problem through Splitting the Unchangeable Text", World Applied Sciences Journal 4 (5): 626-633, 2008.

2) Bharat Singh, Ishadutta Yadav, Suneeta Agarwal, Rajesh Prasad, "An Efficient Word Searching Algorithm through Splitting and Hashing the Offline Text", artcom, pp.387-389, 2009

International Conference on Advances in Recent Technologies in Communication and Computing, 2009.

3) R. J. Enbody and H. C. Du, "Dynamic Hashing Schemes", ACM Computing Surveys, vol. 20, no. 2, 85-113, 1988.
4) P. A. Larson, "Dynamic Hashing", BIT, vol. 18, 184-201, 1978.
5) Sorting and Searching Algorithms, http://www.epaperpress.com
6) Donald E. Knuth, "Sorting and Searching, volume 3 of: The Art of Computer programming", Addison Wesley, 1981.
7) R. S. Boyer, and J. S. Moore, "A fast string-searching algorithm", Communication of ACM, 20(10), pp. 762-772, 1977.
8) A.V.Aho, and M.J. Corasick, " Efficient String Matching: An aid to bibliographic search", Communication of ACM 18(6), pp. 333-340, 1975.
9) R. Prasad and S. Agarwal, "An Efficient String Matching by using Super Alphabet" proc. of the first International Conference on Emerging Trends in Engineering and Technology (available on IEEE Xplore), Nagpur, India., pp. 1181-1186, July 16-18, 2008.

The Design and Implementation of Grid Information Service System Based on Service - Oriented Architecture

Qinghai Bai[1&2]

GJCST Classification
I.3.1, H.3.4

***Abstract*-Grid computing shares the distributed computing resources and unites the virtual organization spread in the different geographic situations to deal with large scale and data intensive computation together. In order to make grid application program use the variety of resources effectively and conveniently some reasonable mechanisms must be adopted to monitor and discover these resources to provide stable, reliable and high-efficiency application environment. This paper main to explore the problem of how to solve the information service application in grid computing environment, construct simulation platform and simulate information service implement procedure by adopting the service-oriented architecture.**

Key words- grid computing; service oriented architecture; grid information service

I. Introduction

Grid concept and technology were initially introduced by Foster and Kesselman in 1998(Ian Foster, Carl Kesselman, 1998). And in 2001, Foster, Kesselman, and Tuecke defined grid as "Coordinated Resource Sharing" that resolves issues in dynamic and multi-organizational virtual structures (Foster, Kesselman, Tuecke, 2001). From then on people conduct comprehensive theoretical research and specific practice on grid computing, drawing high attention from research institutions in plenty of countries and becoming a hot research subject in IT sector.Grid computing introduces coordinated and seamless resource sharing and computing issue. Through grids, grid computing integrates geographically scattered and systematically heterogeneous resources into a virtual "Supercomputer" (Foster, Kesselman. 2004) for largely scaled distributed and high-performance computing. Grid computing offers users huge computing capability by resource sharing and virtualization. It is right its tremendous application potential that IT makes enterprise associations hold greater expectation on grid computing as well.Grid environment is a complex and wide-area distributed system. In grid environment, the computing resources which are in great quantity, heterogeneous and dynamic belong to virtual organizations of different geographic situations. These shared resources and large numbers of users probably result in a series of problems such as the failure of hardware and software, unbalanced load and etc.In order to make the grid users to have access to the variety of resources .effectively and conveniently, many mechanisms must be adopted to monitor and discover the resources to provide the excellent application environment.In the wide-area heterogeneous environment, the resources operating includes the resource organization, discovery, access, location, scheduling, allocation and acknowledgment .The resources in grid include processor, storage resource, directory, network resource, distributed file system, and distributed computing pool and computer cluster. Information service system plays an important role in discovering and monitoring the many resources in different virtual organizations.In grid computing environment, the functions of the information services include the following.To have access to the various static and dynamic information of service system components.To configure and adopt the information services aid to heterogeneous and dynamic environment.With a unified and efficient access information implement interface.Scalability of the access to dynamic information data.To have access to the kind of information resources and distributed-based management.

About[1]. College of Computer Science and Technology, Jilin University,ChangChun, 130000, China

About[2] College of Computer Science and Technology, Inner Mongolia University for Nationalities, Tongliao, 028043, Chinabaiqh68@163.com

II. Service-oriented architecture

1) Overview

SOA is an ideal project to deal with the grid computing environment with application distribution and platform heterogeneity. SOA is a component model which links the application program service by the service definition interfaces. When designed, these interfaces should follow the principle of independence which is independent of service implementation hardware platforms, operating systems and programming languages to construct system services by unified general approach. In SOA, the service which is packed in the business flow is the application program function of the reusable components, which makes the information or business data change from an effective and consistent state to another one. The flow which implements the particular service is not very important. It is enough for the flow to respond the users' command and provide high-quality service for the users' request.By definite communication protocol, mutual manipulation and transparency of the positions can be emphasized through call service. A service appears to be a software component. A service is just like a self-contained function from the point of service requester. Actually, implementation services probably involve many steps carried out by the different computers within the companies or the computers owned by

manybusiness cooperators. As far as software package is concerned, service may be a component or may not be a component. Like class objects, requester application can regard the service as a whole unit. SOA uses service-oriented software packaging technology to provide services by service interfaces and service implementation.SOA includes service description, service discovery and service invocation. Service provider first gives the WSDL description about the service specific implementation and the definite service interfaces, and registers it to the service registry. The service requestor sends out find request, receives the service description from service registry and binds to the service provider by using the service information of the service description to invocate the corresponding service. SOA is a very important model to implement grid information services which can decrease the development difficulty and is convenient to system integration.

2) Web Service

Web service refers to the interfaces which describe some operations (the operations can be accessed by standard XML message transmission mechanism).Web service is described by standard XML concept called the service description of web service. This description includes all the details needed by the service interaction, which are message formats, transport protocols and location. The interface hides the implementation details of the service, allows the programming language used by the writing service of the hardware or software platform which is independent of implementation service and supports the application based on web service to be loosely coupled, component-oriented and cross-technology. Web service fulfills a particular task or a component task, which carries out complex aggregation or commercial transactions independently or together with other web services.Web service belongs to a service based on XML and HTTPS, and it is a new platform for distributed applications which can set up interoperability. The new platform is considered as a set of standards which describe how the application programs implement interoperability on the web. Web service can be written in any language or on any platform, as long as query and access to the service can be implemented by web service standard.The communication protocol is mainly based on SOAP, the service is described by WSDL and metadata car be found and accessed by UDDI.The XML standard used by web service includes the following three parts:SOAP (Simple Object Access Protocol), based on XML,is a kind of message communication protocol within web service application. SOAP defines a XML document format, which describes the method of how to call a remote code.WSDL (Web Service Description Language) is based on XML language, and it is used to describe the web service interfaces.UDDI (Universal Description, Discovery, and Integration) is a industrial standard of service register and discovers on the web service. It defines the SOAP interface to web service registry.

III. MDS4 Information Service

1) The Characteristics of MDS4

MDS supports the construction of VO,which enables the users of VO capable of cooperating and sharing the resources. MDS provides the necessary tools to construct grid information infrastructure based on LDAP.MDS uses LDAP to construct a unified global resource information namespace. MDS adopts attribute-based query mechanism to implement the information query. It supports registry mechanism to renew state information. It also provides the secure access to information. It follows standard GSI andis compatible with the x.509 certificate. The system is extensible and etc.3.2 Architecture of MDS4 MDS4 is a distributed information system, composed of resource layer, aggregation layer and users.

1) Resource layer is made up of one or many service instances which produce service data. These monitoring resources will provide access to resources.
2) Clients, such as user applications, adopt subscription or query request to interact with index service.

2) Service Types of MDS4

MDS4 provides two high layer services, which are index service and trigger service.The index service collects the state information in grid information and publishes it as the resource properties.Trigger service also collects data from resources and meanwhile it monitors the collected data.

IV. Implementation of Grid Information Service

Take the talent exchange grid as an example.To find the real-time resources and services and support the dynamic management to the resources of the virtual organization is the reflection of information service.By adopting service-oriented architecture, it packs accessible grid entities into grid service by web service interfaces to carry out the unified management.

1) Develop Environment and Steps

a) Operation system: ubuntu10.04 windows XP Professional+SP2 in simple Chinese.
b) Toolkits: tomcat U5.0.24: sun java JDK V1.5

The steps are just like the following:

1) create eclipse project
2) add project file
3) Compile and deploy. That is to deploy the grid service to web service container.
4) Service testing and running

2) The Procedure of Implementation

A grid computing environment can be simulated by using six personal computers. The six computers are all installed with Intel Pentium 4. CPUis 2.40 GHZ. The memory is 256M and the hard disk is 320G.They is also installed with integrated sound card and NIC card on main board. Five personal computer (numbered A1, A2, A3, A4 and A5) are installed with ubuntu 10.04 and one (numbered A6) is

integrated sound card and NIC card on main board. Five personal computer (numbered A1, A2, A3, A4 and A5) are installed with ubuntu 10.04 and one (numbered A6) is installed with windows XP professional +SP2 in simple Chinese as a client computer. The network connection speed to your desktop is 100M and 7P-link switch is 100M. Globus toolkits4.2.1 is adopted as grid development platform A1 and A2 are used as resource severs, A3 is used as registry, A6 is used as a client computer and others are not used at present.Server A1 and A2 register separately to registry A3 and they publish to registry server in the form of resource properties. Client A6 enters the system by grid portal and finds the registered resource properties on registry A3 after submitting the task. If A6 finds the resource properties, it will get the access of service interface description service, bind it to the corresponding server and finally pass the result back to client A6 through grid portal.

In the grid computing system of the talent exchange, the" personnel application" is regarded as a grid service. The service provider registers by grid portal and submits resource properties to grid node and the service itself exists on the local server. After the user submits the service request, the system will try to find the service according to the service subscription and will enable the user to get the personnel application by binding the information. There is a rule that only after pre-job training costs are paid off, the user can have access to the service. The pre-job training costs are regarded as a grid service, job application service and pre-job training are assumed to be situated on service A1 and A2, and they are assumed to belong to different management domain. The job application service win have access to the pre-job training costs and subscribe the cost state change .Once users have paid for the training costs, which will cause the change of the cost state, the job-application service will get the notification to enable it to get the cost state. And this is called subscription or notification mechanism. By subscription or notification mechanism, when service state begins to change, the subscriber will be notified timely, and the subscriber will take the corresponding action according to the trigger event to make sure that the users will get job application service.

V. Conclusions

Service publishing, discovery and binding operation are carried out by adopting service-oriented architecture which is concerned on web service, composed of three basic elements of service description, service discovery and service invocation and plays the three roles as service providers, service consumers and service registry. Some main problems involved in some grid information services are solved by simulating a grid service implementation of information services which takes MDS4, the component of Globus toolkits 4.2.1, as information service tool. By the service interfaces of web service, grid entities are packed into grid service and the concrete realization of grid entities is shielded, which appears to be a unified interface to meetthe clients' concept of invocating the service on demand.This project is supported by National Natural Science Foundation of China (No. 60873235&60473099) and by Science-Technology Development Key Project of Jilin Province of China (No. 20080318) and by Program of New Century Excellent Talents in University of China (No. NCET-06-0300).

VI. References

1) Ian Foster, Carl Kesselman (Eds).The Grid: Blueprint for a New Computing Infrastructure (1st edition). Morgan Kaufmann publishers, San Francisco, USA (1 November 1998).
2) Foster, I., Kesselman, C., Tuecke, S.The Anatomy of the Grid: Enabling Scalable Virtual organizations. International Journal of Supercomputer Applications, 15(3), 2001.
3) I. Foster, C. Kesselman.The Grid 2: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, 2004.
4) United States]Mao Zhen Li,Mark Baker, Ed , WANG xianglin,ZHANG shanqing,WANGjingli, Translation. Grid Computing Core Technology[M]. Tsinghua University Press, Beijing, China, 2006.12, 105-107.
5) GMA White paper,http://www-didc.lbl.gov/GGF-PERF/GMA-WG/.
6) United States]Joshy Joseph, Craig Fellenstein ,Ed, ZHAN xiaosu,ZHANG shaohua Translation. Grid Computing [M]. Tsinghua University Press, Beijing, China, 2005.1, 58-62.
7) MA hengtai,LI pengfei,YAN xuexiong etc.Web Services security[M].Electronic Industry Press , Beijing ,China,2007.12,101-105.

Performance Analysis of the Postcomputation-Based Generic-Point Parallel Scalar Multiplication Method

Turki F. Al-Somani

GJCST Classification
F.3.1, F.1.2

***Abstract*- A Postcomputation-based generic-point parallel scalar multiplication method has recently been proposed for high-performance end servers that employ parallel elliptic curve cryptoprocessors. The sequential precomputation overheads, in the postcomputation-based method, are replaced with parallelizable postcomputations. This paper analyzes the performance of the postcomputation-based method with 128 ≤ m ≤ 256 using a number of parallel elliptic curve cryptoprocessors. The results show that the best performance is achieved when eight cryptoprocessors are used.**

Keywords- Elliptic Curve Cryptosystems, Parallel Scalar Multiplication, Precomputations, Postcomputations, Generic-Point Scalar Multiplication.

I. Introduction

Elliptic curve cryptosystems (ECCs), which were initially proposed by Niel Koblitz and Victor Miller in 1985 [1], are seen as serious alternatives to the RSA system but with a much shorter word length. An ECC with a key size of 128 – 256 bits has been shown to offer equal security to an RSA system with a key size of 1 – 2 Kbits [2]. To date, no significant breakthroughs have been made in determining the weaknesses of ECCs, which are based on a discrete logarithm problem over points on an elliptic curve. The fact that the problem appears so difficult to crack means that key sizes can be considerably, and possibly even exponentially, reduced [3]. This advantage of ECCs has gained recognition recently, and has resulted in their incorporation in many standards such as IEEE, ANSI, NIST, SEC and WTLS.Scalar multiplication is the basic operation for ECCs. Scalar multiplication of a group of points on an elliptic curve is analogous to the exponentiation of a multiplicative group of integers modulo a fixed integer m. The scalar multiplication operation, denoted as kP, where k is an integer and P is a point on the elliptic curve, represents the addition of k copies of point P. Scalar multiplication is then computed by a series of point doubling and point addition operations of the point P that depends on the bit sequence that represents the scalar multiplier k. Several scalar multiplication methods have been proposed [4]. However, for high-performance end servers, the current sequential scalar multiplication methods are too slow to meet the demands of increasing number of customers.

About-Turki F. Al-Somani is with the Department of Computer Engineering, Faculty of Engineering, Al-Baha University, Al-Baha, Saudi Arabia.Email: tfsomani@bu.edu.sa

Identifying efficient scalar multiplication methods for such servers has thus become crucial. Scalar multiplication methods that can be parallelized are often used for high-speed implementations. Precomputations [4-6] have also been applied to speed up scalar multiplication, but require sequential steps that cannot be parallelized, and are primarily advantageous basically when the elliptic curve point is fixed. However, during secure communication sessions that use public keys the elliptic curve point changes, as it depends on the public key of the communicating entity, that is, it is session dependant. This is also the case when digital signatures are used. Hence, the computation of scalar multiplications is generally performed with a generic elliptic curve point. Because the elliptic curve point is likely to differ in each session, the overheads resulting from the necessary precomputations must be considered when estimating the total computational time required. Postcomputations have recently been proposed [7] as an alternative method to speedup scalar multiplications. In [7], the precomputation overheads are replaced by postcomputations that can be parallelized. This paper shows that the concurrent precomputation of several points outperforms the method proposed in [7] with the same number of points and parallel processors. The remainder of the paper is organized as follows. Section 2 introduces the basic ECC. Section 3 describes the postcomputation-based generic-point parallel scalar multiplication method [7]. Section 4 presents a performance analysis of the postcomputation-based method. Section 5 concludes.

II. Elliptic Curve Crypto Preliminaries

Elliptic curve cryptosystems (ECCs) [4] have attracted much research attention and have been included in many standards. ECCs are evolving as an attractive alternative to other public-key schemes such as RSA by offering a smaller key size and a higher strength per bit. Extensive research has been conducted on the underlying math, security strength and efficient implementations of ECCs. Of the various fields that can underlie elliptic curves, prime fields $GF(p)$ and binary fields $GF(2^m)$ have proved to be best suited to cryptographic applications. An elliptic curve E over the finite field $GF(p)$ defined by the parameters $a, b \in GF(p)$ with $p > 3$ consists of the set of points $P = (x, y)$, where $x, y \in GF(p)$, that satisfies the equation

$$y^2 = x^3 + ax + b \quad (1)$$

where $a, b \in GF(p)$ and $4a^3 + 27b^2 \neq 0 \bmod p$, together with the additive identity of the group point O known as the

„point at infinity' [4]. The number of points #*E* on an elliptic curve over a finite field *GF*(*q*) is defined by Hasse's theorem [4]. The discrete points on an elliptic curve form an abelian group, the group operation of which is known as „point addition'. Elliptic curve point addition is defined according to the „chord-tangent process'. Point addition over *GF*(*p*) can be described as follows.Let *P* and *Q* be two distinct points on *E* defined over *GF*(*p*) with $Q \neq P$ (*Q* is not the additive inverse of *P*). The addition of the points *P* and *Q* gives the point *R* ($R = P + Q$), where *R* is the additive inverse of *S* and *S* is a third point on *E* intercepted by the straight line through points *P* and *Q*. The additive inverse of point $P = (x, y) \in E$ over *GF*(*p*) is the point $-P = (x, -y)$, which is the mirror of point *P* with respect to the *x-axis* on *E*. When $P = Q$ and $P \neq -P$, the addition of *P* and *Q* is the point R ($R = 2P$), where *R* is the additive inverse of *S* and *S* is the third point on *E* intercepted by the straight line tangential to the curve at point *P*. This operation is referred to as point doubling. The finite field $GF(2^m)$ has particular importance in cryptography, as it leads to very efficient hardware implementations. Elements of the field are represented in terms of a basis. Most implementations use either a polynomial basis or a normal basis [8]. Letting $GF(2^m)$ be a finite field of characteristic two, a non-supersingular elliptic curve *E* over $GF(2^m)$ can be defined as the set of solutions $(x, y) \in GF(2^m) \times GF(2^m)$ to the equation

$$y^2 + xy = x^3 + ax^2 + b \tag{2}$$

where *a* and $b \in GF(2^m)$, $b \neq 0$, together with the point at infinity. It is well known that *E* forms a commutative finite group, with *O* as the group identity, under the addition operation known as the tangent and chord method. Explicit rational formulas for the addition rule involve several field arithmetic operations (addition, squaring, multiplication and inversion) in the underlying finite field. The group operations in affine coordinate systems involve finite field inversion, which is a very costly operation, particularly for prime fields. Projective coordinate systems can be used to eliminate the need to perform inversions. Several projective coordinate systems have been proposed in the literature, including the homogeneous, Jacobian, Chudnovsky-Jacobian, modified Jacobian, Lopez-Dahab, Edwards and mixed coordinate systems [9][10].Several scalar multiplication methods have been proposed in the literature [4]. Computing *kP* can be achieved with a straightforward binary method – the so-called double-and-add method – based on the binary expression of the multiplier *k*. *kP* can be computed using a binary method as follows.

Let $k = (k_{m-1},\ldots,k_0)$, where k_{m-1} is the most significant bit of *k*, be the binary representation of *k*. The multiplier *k* can be written as

$$k = \sum_{0 \le i < m} k_i 2^i = k_{m-1}2^{m-1} + \cdots + k_1 2 + k_0 \tag{3}$$

Using the Horner expansion, *k* can be rewritten as

$$k = (\cdots((k_{m-1}2 + k_{m-2})2 + \cdots + k_1)2 + k_0) \tag{4}$$

Accordingly,

$$kP = 2(\cdots 2(2k_{m-1}P + k_{m-2}P) + \cdots + k_1 P) + k_0 P \tag{5}$$

The algorithm for the binary method is as follows.

Algorithm 1: Binary Method

(1) Input *P*, *k*.
(2) $Q \leftarrow O$.
(3) For *i* from *m* – 1 down to 0, perform
a. $Q \leftarrow 2Q$,
b. If $k_i = 1$, then $Q \leftarrow Q + P$.
(4) End for.
(5) Output *Q*.

(6)

The binary scalar multiplication method is the most straightforward scalar multiplication method. It inspects the bits of the scalar multiplier *k*. If the inspected bit $k_i = 0$, then only point doubling is performed. If, however, the inspected bit $k_i = 1$, then both point doubling and addition are performed. The binary method requires *m* point doublings and an average of *m*/2 point additions. Non-adjacent form (NAF) reduces the average number of point additions to *m*/3 [11]. With NAF, signed-digit representations are used such that the scalar multiplier's coefficient $k_i \in \{0, \pm 1\}$. NAF has the property that no two consecutive coefficients are nonzero. It also has the property that every positive integer *k* has a unique NAF encoding, which is denoted as NAF(*k*).

III. The Postcomputation-based Method

The essential concept underlying the method proposed in [7] is the replacement of sequential precomputations with parallelizable postcomputations. Multiplier k in [7] is partitioned into u partitions that can be processed in parallel by u processors using the binary method. Some of the postcomputations are then distributed on $u - 1$ processors to be performed in parallel. The points that result from processing these key partitions with the postcomputations are then assimilated to produce kP.

Let $k = (k_{m-1}, \ldots, k_0)$, where k_{m-1} is the most significant bit of k, be the binary representation of multiplier k. Then, after partitioning k into u partitions, multiplier k can be written as

$$k = (k^{(u-1)} || k^{(u-2)} || \ \ldots || k^{(0)}) \tag{6}$$

Scalar multiplication product kP can then be computed as

$$kP = \sum_{0 \le i \le u} t_i \tag{7}$$

where t_i is defined as

$$t_i = 2(\cdots 2(2k_{iv+v-1}(2^{iv}P) + k_{iv+v-2}(2^{iv}P)) + \cdots + k_{iv+1}(2^{iv}P)) + k_{iv+0}(2^{iv}P) \tag{8}$$

A key observation is that Eq. (8) can be rewritten as

$$t_i = (2^{iv})[2(\cdots 2(2k_{iv+v-1}P + k_{iv+v-2}P) + \cdots + k_{iv+1}P) + k_{iv+0}P] \quad (9)$$

Eq. (9) implies that the required precomputations of Eq. (8) can be replaced by postcomputations, which are point doublings. Each partition requires iv point doublings to produce the correct partial product. To balance the number of point operations, we need to balance the total number of field multiplications, as field multiplication is the dominant type of operation in elliptic curve point operations in projective coordinates. This implies that multiplier k should be partitioned into u partitions of different sizes, as shown in Eq. (10).

$$m = m_{(u-1)} + m_{(u-2)} + \cdots + m_{(1)} + m_{(0)} \quad (10)$$

Accordingly, the number of bits in partition $t_{(i)}$ must be greater than the number of those in $t_{(i+1)}$ and fewer than the number of those in $t_{(i-1)}$, as can be seen from Eq. (11).

$$m_{(u-1)} < m_{(u-2)} < \cdots < m_{(1)} < m_{(0)} \quad (11)$$

Assume that the double and add point operations require r and s field multiplications, respectively. Then, let the total number of field multiplications in partition $k^{(i)}$ equal $M_{(i)}$. Because partition $k^{(0)}$ is the only one to require no postcomputations, a balanced number of point operations can be reached by solving Eqs. (10) and (11) together with the following equations (12-14).

$$M_{(0)} = m_{(0)}(r) + \frac{m_{(0)}}{2}(s) \quad (12)$$

$$M_{(i)} = m_{(i)}(r) + \frac{m_{(i)}}{2}(s) + (r)\sum_{0 \le j < i} m_j \quad (13)$$

$$M_{(0)} = M_{(1)} = \cdots = M_{(u-1)} \quad (14)$$

The computation of kP in parallel without precomputations can be performed efficiently using the following algorithm.

Algorithm 2: Postcomputation-based Method

1. Inputs: P, k
2. By padding k with zeros if necessary, solve Eqs. (10)-(14) together, and write $k = (k^{(u-1)}||k^{(u-2)}|| \ldots ||k^{(0)})$, where $k^{(i)}$ is a partition of length $m_{(i)}$ bits.
3. Initialisation: $Q \leftarrow P$, $R \leftarrow O$.
4. Parallel Scalar Multiplication:
   - 4.1. For $i = 0$ to $u - 1$ do in parallel
     - 4.1.1. $Q \leftarrow$ Binary method $(k^{(i)}, P_i)$
     - 4.1.2. If $(i > 0)$, then
       - 4.1.2.1. for $c = 0$ to $((\sum_{0 \le j < i} m_j) - 1)$ do
         - 4.1.2.1.1. $Q \leftarrow 2Q$
     - 4.1.3. $R \leftarrow R + Q$
5. Output R

Example: Let $k = (1000\ 0101\ 1100\ 0011)_2 = (34243)_{10}$, $m = 16$, $u = 4$ and $r = \frac{s}{2}$. The sizes of the key partitions are $m_0 = 9$, $m_1 = 4$, $m_2 = 2$ and $m_3 = 1$. The key partitions are $k^{(0)} = 111000011$, $k^{(1)} = 0010$, $k^{(2)} = 00$, and $k^{(3)} = 1$. The scalar multiplication of these partitions is then computed in parallel according to the following.

$t_0 = 2(2(2(2(2(2(2(1)P + (1)P) + (1)P) + (0)P) + 0P+0P+0P+1P+1P=451P,$

$t_1 = (2^9)[2(2(2(0)P + (0)P) + (1)P) + (0)P] = 1024P,$

$t_2 = (2^{13})[2(0)P + (0)P] = O$

$t_3 = (2^{15})[(1)P] = 32768P.$

Finally, kP is computed as

$kP = t_0 + t_1 + t_2 + t_3 = 451P + 1024P + O + 32768P = 34243P$ ■

IV. Performance Analysis

The time complexity of the proposed method in [7] equal to $(m_{(0)})$ point doublings + $(\frac{m_{(0)}}{2} + log_2(u))$ point additions and $(m_{(0)})$ point doublings + $(\frac{m_{(0)}}{3} + log_2(u))$ point additions using binary and NAF encoding, respectively. However, the proposed method in [7] has not been analyzed when different values of m and u are used. Table 1 shows the lengths, in number of bits, of each key partition. Table 1 also shows that no more than 8 processors should be used with the proposed method in [7]. This is clearly shown when 12 processors are used and only 8 of the processors are utilized.

Table 2 shows the results for the method proposed in [7]. In Table 2, the first two columns show the key size m and the number of parallel processors u. The third column shows the length of the first key partition $m_{(0)}$. The number of point doublings, additions and accumulation additions are shown in the following columns. It is assumed here that the required computation time for point addition is twice that required for point doubling. Accordingly, columns 8 and 9 show the total number of point doublings and additions for the point doublings for binary and NAF encoding, respectively. The results of Table 2 are depicted in Figure 1 and 2 for binary and NAF encoding, respectively. Clearly, the results show that the best performance is achieved when eight processors are used. Increasing the number of processors, however, does not mean better performance.

V. Conclusion

Sequential scalar multiplication methods are too slow for high-performance end servers because of the demand resulting from increasing numbers of customers. Existing parallel methods, however, require sequential precomputations for each new session. Recently, the first generic-point parallel scalar multiplication method has been proposed. In the proposed method, the precomputation overhead is replaced by postcomputations that can be

parallelized. In this paper we have analyzed the performance of the postcomputation-based method with $128 \leq m \leq 256$ using a number of parallel elliptic curve cryptoprocessors. The results show that the best performance is achieved when eight cryptoprocessors are used.

VI. References

1) Koblitz, N. (1987) Elliptic curve cryptosystems, *Mathematics of Computation*, 48, 203-209.
2) Rivest, R., Shamir, A. and Adleman, L. (1978) A method for obtaining digital signatures and public key cryptosystems, *Communications of the ACM*, 21, 2, 120-126.
3) Blake, I., Seroussi, G. and Smart, N. (1999) *Elliptic Curves in Cryptography,* Cambridge University Press, New York.
4) Hankerson, D., Menezes, A. J. and Vanstone, S. (2004), *Guide to Elliptic Curve Cryptography*, Springer-Verlag.
5) Brickell, E. F., Gordon, D. M., McCurley, K. S. and Wilson, D. B. (1993) Fast exponentiation with precomputation, *Advances in Cryptology – Eurocrypt'92,* LNCS 658, pp. 200-207. Springer-Verlag.
6) Lim, C. H. and Lee, P. J. (1994) More flexible exponentiation with precomputations. *Proc. CRYPTO 94*, pp. 95-107.
7) Al-Somani, T. F. and Ibrahim, M. K. (2009) Generic-point parallel scalar multiplication without precomputations, *IEICE Electronics Express*, 6, 24, 1732-1736.
8) Lidl, R. and Niederreiter, H. (1994) *Introduction to Finite Fields and their Applications.* Cambridge University Press, Cambridge, UK.
9) Cohen, H., Ono, T. and Miyaji, A. (1998) Efficient elliptic curve exponentiation using mixed coordinates. *Advances in Cryptology – SIACRYPT '98,* LNCS 1514, pp. 51-65. Springer-Verlag.
10) Washington, L. C. (2008), *Elliptic Curves: Number Theory and Cryptography, 2nd ed.*, CRC Press.
11) Joye, M. and Tymen, C. (2001) Compact encoding of non-adjacent forms with applications to elliptic curve cryptography, *Public Key Cryptography, LNCS 1992*, pp. 353-364. Springer-Verlag.

Table 1: Results of the method proposed in [7] with $m = 128, 160, 200, 256$ and $u = 2, 4, 8, 12$.

m	u	m_0	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}
128	2	85	43	-	-	-	-	-	-	-	-	-	-
	4	68	34	17	9	-	-	-	-	-	-	-	-
	8	64	32	16	8	4	2	1	1	-	-	-	-
	12	64	32	16	8	4	2	1	1	0	0	0	0
160	2	107	53	-	-	-	-	-	-	-	-	-	-
	4	85	43	21	11	-	-	-	-	-	-	-	-
	8	80	40	20	10	5	3	1	1	-	-	-	-
	12	80	40	20	10	5	3	1	1	0	0	0	0
200	2	133	67	-	-	-	-	-	-	-	-	-	-
	4	107	53	27	13	-	-	-	-	-	-	-	-
	8	100	50	25	13	6	3	2	1	-	-	-	-
	12	100	50	25	13	6	3	2	1	0	0	0	0
256	2	171	85	-	-	-	-	-	-	-	-	-	-
	4	137	68	34	17	-	-	-	-	-	-	-	-
	8	129	64	32	16	8	4	2	1	-	-	-	-
	12	128	64	32	16	8	4	2	1	1	0	0	0

Table 2: Results of the method proposed in [7] with $m = 128, 160, 200, 256$ and $u = 2, 4, 8, 12$.

m	u	m_0	DBLs	Binary ADDs	NAF ADDs	Acc ADDs	Binary Total (DBLs)	NAF Total (DBLs)
128	2	85	85	43	28	1	173	143
	4	68	68	34	23	2	140	118
	8	64	64	32	21	3	134	112
	12	64	64	32	21	4	136	114
160	2	107	106	53	35	1	214	178
	4	85	85	43	28	2	175	145
	8	80	80	40	27	3	166	140
	12	80	80	40	27	4	168	142
200	2	133	133	67	44	1	269	223
	4	107	107	54	36	2	219	183
	8	100	100	50	33	3	206	172
	12	100	100	50	33	4	208	174
256	2	171	171	86	57	1	345	287
	4	137	137	69	46	2	279	233
	8	129	129	65	43	3	265	221
	12	128	128	64	43	4	264	222

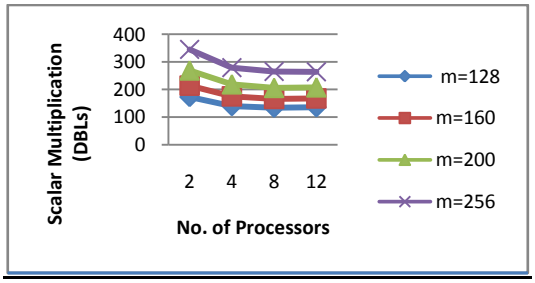


Figure 1: Binary encoding results of the method proposed in [7] with $m = 128, 160, 200, 256$ and $u = 2, 4, 8, 12$.

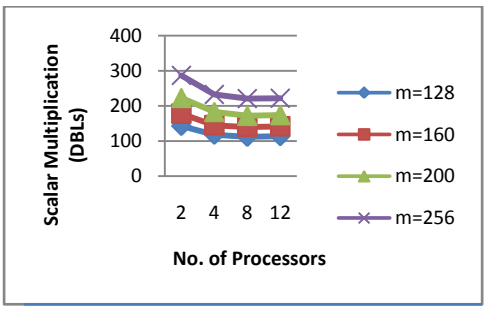


Figure 2: NAF encoding results of the method proposed in [7] with $m = 128, 160, 200, 256$ and $u = 2, 4, 8, 12$.

22) S.Ganapathy and S.Velusami," Design of MOEA based Decentralized Load-Frequency Controllers for Interconnected Power Systems with AC-DC Parallel Tie-lines", International Journal of Recent Trends in Engineering, Vol .2, No. 5, .p:357-361, November 2009.
23) Goldberg, "Genetic Algorithms in search, optimization and machine learning" Addison-Wesly, 1989
24) Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali and Osman A. Sadek, "Using Genetic Algorithm to Improve Information Retrieval Systems", World Academy of Science and Engineering Technology, Vol.17, No.2,pp.6-13, May 2006
25) Zheng and Kiyooka, "Genetic Algorithms Applications: Assignment #2 for Dr. Z. Dong," November 1999
26) Tomoyuki Hiroyasu, "Diesel Engine Design using Multi-Objective Genetic Algorithm", Technical Report, 2004
27) Hazra, J. Phulpin, Y. Ernst, D., "HVDC control strategies to improve transient stability in interconnected power systems", Power Tech, 2009, IEEE Bucharest p.p. 1 - 6 DOI: 10.1109/PTC.2009.5281816

A2Z Control System- DTMF Control System

Er. Zatin Gupta[1], Payal Jain[2] , Monika[3]

GJCST Classification
H.4.3

***Abstract*-Dual Tone Multi Frequency (DTMF) technique for controlling the domestic and industrial appliances is being presented in this paper. A simple mobile phone which works on DTMF tone, used to control the domestic as well as industrial electrical appliances which with the control system which we have designed here for experimental study. In recent state of affairs, domestic, military and industrial applications use this technique because it can be operated from remote location. Radio frequency (RF) is also used for wireless communication but DTMF is an alternate for RF. Mobile phone is used to send the DTMF code from remote location to the control system. The blocks of system are mobile phone, Microcontroller (AT89S52), DTMF Decoder (MT8870D), Relays and power supply. This paper shows the working areas where the system is applicable and how it has advantages over RF.**

I. Introduction

The main features of DTMF Decoder (MT8870D) are as follows:• DTMF Receiver
• Low power consumption
• Power-down mode
• Inhibit mode
Main applications are as follows:
• Receiver system
• Remote control
• Telephone answering machine

1) BLOCK DIAGRAM

The block diagram of system consists of the following equipments:

a) Mobile Phone

Wireless control of domestic and industrial appliances is the objective of this paper so to achieve the objective, a mobile is used here. Particular button of mobile keypad produces specific DTMF tone that will transmit by the operator of mobile to the control system.

b) Dtmf Decoder (Mt8870d)

The MT8870D is DTMF receiver consisting of digital decoder. It uses digital counting techniques for detecting and decoding all 16 DTMF tone pairs into a 4 bit- code. The micro controller takes the 4 bit code as input.

About[1]. Lecturer, Department of Information and Technology, Maharishi Markandeshwar Engineering College, Maharishi Markandeshwar University, Mullana, Ambala (Haryana) zatin.gupta2000@gmail.com, Mobile No: +919050406988
About[2]. Lecturer, Department of Information and Technology, Maharishi Markandeshwar Engineering College, Maharishi Markandeshwar University, Mullana, Ambala (Haryana) payaljain2006@gmail.com, Mobile No: +919466742552.
About[3]. Lecturer, Department of Information and Technology, Maharishi Markandeshwar Engineering College, Maharishi Markandeshwar University, Mullana, Ambala (Haryana) monawadhwa@gmail.com

c) Microcontroller (At89s52)

Microcontroller is the control unit of this system. We have used AT89S52. It is low power, high performance CMOS 8-Bit controller with 4K bytes of ROM and 128 bytes of RAM.

d) Dtmf Signal

DTMF is most widely known method of Multi Frequency Shift Keying (MSFK) data transmission technique. DTMF was developed by Bell Labs to be used in the telephone system. Most telephones today uses DTMF dialing (or "tone" dialing).

	1209 Hz	1336 Hz	1477 Hz	1633 Hz
697 Hz	1	2	3	A
770 Hz	4	5	6	B
852 Hz	7	8	9	C
941 Hz	*	0	#	D

Table 1: DTMF signal frequency encoding table.

The DTMF technique outputs distinct representation of 16 common alphanumeric characters (0-9, A-D, *, #) on the keypad. The lowest frequency used is 697Hz and the highest frequency used is 1633Hz, as shown in Table 1.

e) Relay Logic

To ON/OFF the appliances we need relays because relay is an on/off switch that can control by microprocessor automatically. Appliances are connected to these relays. Relays are of two types SPDT, DPDT, here SPDT Relays used.

2) EXPERIMENTAL STUDY

The main controlling technique using DTMF can be explained as follow:Integrated DTMF Receiver

a) Features

It has low power consumption, adjustable guard time, inhibit mode. These are the features, which we have used here to achieve the goal of this paper.

b) Applications

It is widely used in Mobiles, Remote control, Personal computers, Telephone answering Machine, etc.

II. Explanation of MT8870D

The MT8870D is a complete DTMF receiver integrating both the band splitFilter and digital decoder functions. The filter section uses switched capacitor techniques for high and low group filters; the decoder uses digital counting techniques to

detect and decode all 16 DTMF tone pairs into a 4-bit BCD code.

1) PRACTICAL DESCRIPTION

The MT8870D monolithic DTMF receiver has small size, low power consumption and high performance. Its architecture consists of a band split filter section, which separates the high and low frequency tones, followed by a digital counting section which verifies the frequency and duration of the received tones before passing the corresponding code to the output bus to the Micro Controller.

a) Filter Section

Separation of the low and high group tones is achieved by applying the DTMF signal to the inputs of band pass filters, the bandwidths of which correspond to the low and high group frequencies.

b) Decoder Section

Decoder Section having the digital counting techniques to determine the frequencies of the incoming tones and to verify that they correspondingly generate a standard DTMF frequency.

c) Steering Circuit

Before registration of a decoded tone pair, the receiver checks for a valid signal duration. This check is performed by an external RC time constant.

d) Power-Down Feature

+ 5V voltage applied to pin no 6 (PWDN) to power down the device which minimize the power consumption in a standby mode. It stops the oscillator and the functions of the filters.

e) Inhibit Mode

Inhibit mode is enabled by applying +VE voltage to the pin 5 (INH). It inhibits the detection of tones representing characters A, B, C, and D. The output code will remain the same as previously received.

2) ULN 2003A

The ULN2003A is a high voltage; high current Darlington pair arrays each containing seven open collector Darlington pairs with common emitters. Each channel rated at 500mA and can withstand peak currents of 600mA. Suppression diodes are included for inductive load driving and the inputs are pinned opposite the outputs to simplify board layout and it is used to operate relays.

The main features of ULN 2003A are as
Follows:
a) Seven Darlington pair Per Package
b) Output Current 500mA per Driver
c) Output Voltage 50V
d) Integrated Suppression Diodes for Inductive loads.
e) Outputs can be paralleled for higher current.

III. EXPERIMENTAL STUDY

The experimental study of DTMF communication is to be carried out in three stages. The FIRST STAGE is testing the output of mobile phone or DTMF Encoder (91241B). The SECOND STAGE is decoding of transmitted data. The THIRD STAGE is to perform the operation which is supposed to be done by pressing the particular key of mobile phone keypad.

1) First Stage

The goal of this stage is to verify that the DTMF signal produced and transferred to the control system is appropriate and not affected by noise. The control System with DTMF receiver is programmed to perform a series of operations. The DTMF characters can be heard as “beeps”. We make such an arrangement of taking the DTMF tone and then checks its frequency with the standard and then matches both the results if they are same then it means that are signal is not distorted and hence supplied to the control system.

2) Second Stage

This stage is carried out to decode the DTMF signal and to produce the output which is sent to the Micro-Controller. The decoding is done by the DTMF Decoder (MT8870D) and it gives the output in BCD format i.e. The FOUR digit code which is sent to the Micro-Controller. In this control system we are presenting that code also on LED’s so that we can easily decode the signal and one LED is used to tell that a DTMF signal is received and decoded i.e. whenever a DTMF signal is received and decoded by decoder then a signal LED will blink. By this means we decode the DTMF Signal.

3) Third Stage

This stage is to perform the operation which is supposed to be done by the decoding the DTMF signal into a particular BCD code that represents a decimal digit basically. To do this, an experiment is performed using distinct values and repeated it many times to test the onboard DTMF receiver. The commands, as illustrated in Table are directly represented by the DTMF characters. Since the main focus is to check that whether the operations defined in coding of the system are working properly, as they actually supposed to be or not. If there is any problem then it means that an error occurs there in between so again the process started from SECOND STAGE until we got the appropriate result.

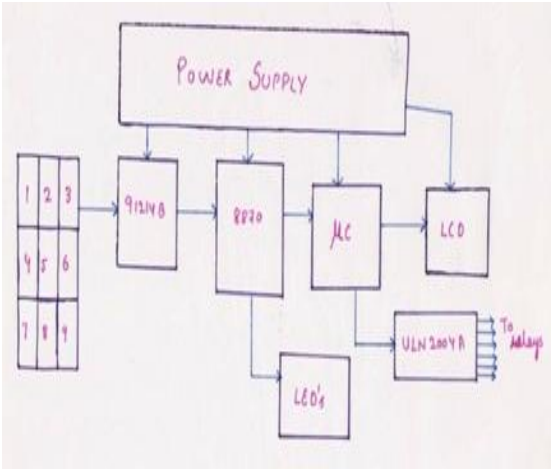
X	0	1	2	3	4	5
Y	Off All	On First	On Second	Off First	Off Second	On All

Table shows the operations performed

X tells DTMF signal produced by such key & Y tells Operation performed. To check the working of system, we gave values in between 0-5 to the system and test the functionality of the control system.

IV. A2Z CONTROL SYSTEM

Block Diagram

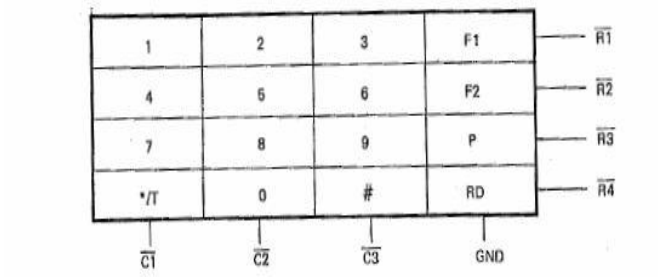


Working Image

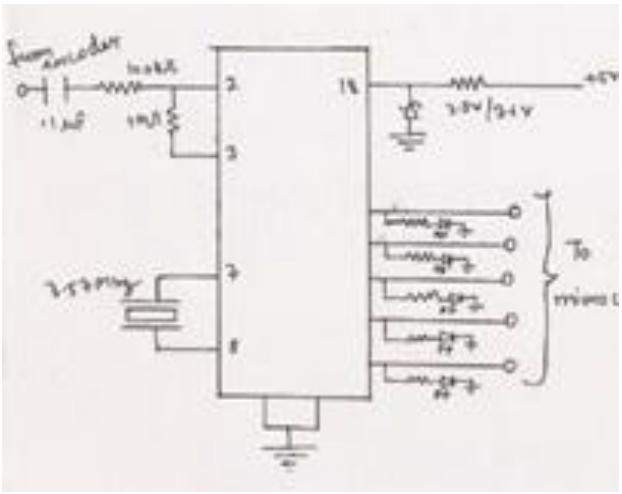
Our control system has two relays to operate two devices such as to charge mobile phone and to blow the bulb.

Keypad Assignment

Keyboard Assignment



Decoder Circuit



V. CONCLUSION

This paper has described the design and specifications of the DTMF based control system and it is feasible to make such a control system also it has many advantages over any other media i.e. RF because RF can be easily received by any other antenna but DTMF tone cannot be. As we are making a call from a mobile to the mobile which is connected to the system then only that particular mobile will receives the signal and decode it. This approach is very secure and is very useful in domestic as well as industrial usage. Although DTMF has advantages like low cost, simplicity, very popular in telephone industry so we conclude that this kind of control systems are very useful in today's scenario.

VI. REFERENCES

To complete this paper successfully we need to refer with many websites and papers, some of them are as mentioned below:

1) A. Shatnawi, A. Abu-El-Haija, A. Elabdalla, "*A Digital Receiver for Dual-Tone Multi-frequency (DTMF) Signals*", Technology Conference, Ottawa, CA, May 1997.
2) M. Felder, J. Mason, B. Evans, "*Efficient Dual-Tone Multi-frequency Detection Using Non-uniform Discrete Fourier Transform*", IEEE Signal Processing Letters, Vol. 5, No. 7, July 1998.
3) A. Plaisant, "*Long Range Acoustic Communications*", OCEANS '98 Conference Proceedings, Vol. 1, pp. 472, Sept. 1998.
4) Z. Wei-Qing, W. Chang-Hong, P. Feng, Z. Min, W. Rui, Z. Xiang-Jun, D. Yong-Mei, "*Underwater Acoustic Communication System of AUV*", OCEANS '98 Conference Proceedings, Vol. 1, pp. 477, Sept. 1998.
5) M. J. Callahan Jr., "*Integrated DTMF Receiver*", IEEE J. Solid States Circuits, vol. SC-14, pp. 85-90, Feb. 1979.
6) www.alldatasheet.com

7) www.active-robots.com/.../av-modules.shtml - United Kingdom
8) Robots, Androids, and Animations 12 Incredible Projects You Can Build" by John Iovine Second Edition McGraw-Hill.
9) PDA Robotics Using Your Personal Digital Assistant to Control Your Robot" by Douglas H. Williams McGraw-Hill
10) Remote robot control system based on DTMF of mobile phone" published in 6th IEEE International Conference.

A Proposal for a Biometric Key Dependent Cryptosystem

GJCST Classification
D.4.6

K. Hassanain[1] ,M. Shaarawy, E. Hesham[2]

***Abstract*-With the increasing reliance on electronic information, which needs to be exchanged across the internet or stored on open networks, cryptography is becoming an increasingly important feature of computer security. A biometric key dependent cryptosystem is proposed, to ensure the security of the whole system by using fingerprint features as a key in a cryptosystem, like, key-dependent Advanced Encryption Standard (KAES). KAES is used to ensure that no trapdoor is present in cipher and to expand the key-space to slow down attacks.**

Keywords- AES, KAES, MD5, RNG, PRNG, SHA-1.

I. Introduction

Cryptography is becoming an increasingly important feature for information security, and there are many available cryptographic algorithms for securing information: Symmetric and Asymmetric. The strength of cryptosystem depends on many factors: key length, algorithm complexity and resistance to cryptanalysis techniques [1][2].There are mainly two problems when using traditional password or token as a key for any cryptosystem. First, the security of the key, and hence the cryptosystem, is now only as good as the password. Due to practical problems of remembering various passwords, some users tend to choose simple words, phrases, or easily remembered personal data, while others resort to write the password down on an accessible document to avoid data loss. The second problem is the lack of a direct connection between the password and the user, as a password is not tied to a user, a system running the cryptographic algorithm is unable to differentiate between the legitimate user and an attacker who fraudulently acquires the password of a legitimate user (Authentication) [1].An alternative to password protection, there are many approaches to bind a crypto-key with biometrics. The famous two approaches are a biometric-based key release and biometric-based key generation [2][3]. In biometric-based key release the key is hidden into a biometric template at the enrollment phase and is available to be released at authentication phase. While the other, the key is generated directly from the biometric data using one of secure hash functions [4].This paper introduces a biometric cryptosystem in which the key is generated from biometric data and produced key is used in key dependent encryption algorithm to ensure the security of the system and slow down its attacks. The paper is organized as follows: Section II presents the proposed biometric key dependent cryptosystem. Section III explains the evaluation criteria. Section IV discusses theexperimental results. Section V summaries and concludes the paper. References are given in Section VI.

About[1]- Faculty of computers and Information, Helwan University, Egypt. (email: mhmdshaarawy@yahoo.com, Hesham.eman@gmail.com
About[2]- Technical Research Department, Egypt(email: khass@idsc.net.eg

II. A Biometric Key Dependent Cryptosystem

The proposed scheme replaces the secret key in a cryptosystem with a key which generated directly from one of the human biometric data (e.g. fingerprint). In general, the proposed biometric-key cryptosystem could be subdivided into three phases: biometric phase, key generation phase and encryption phase. Figure 1 shows the overall structure of the proposed system.

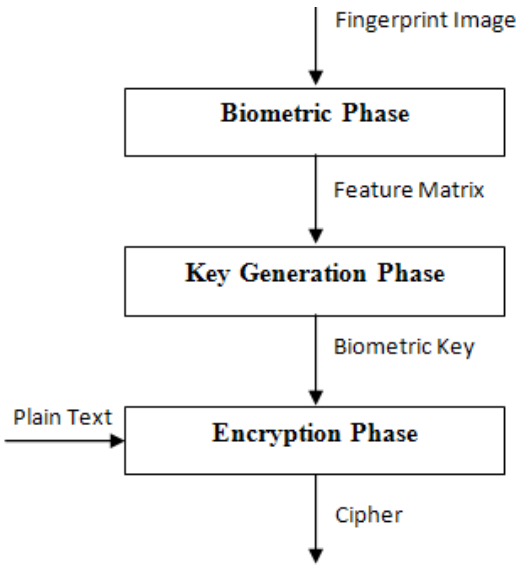


Fig.1. The proposed biometric key dependent cryptosystem

In the proposed system, the input to the biometric phase is fingerprint image which acquired from the system user's finger using fingerprint reader. Through this phase some unique characteristics of the fingerprint image are extracted to form a biometric feature matrix.The produced matrix is used as an input to the next phase to generate a 128-bit key using one of cryptographic hash functions such as Secure Hash Algorithm (SHA-1) or Message-Digest algorithm 5 (MD5). The plain-text is then encrypted using the generated key by one of cryptographic encryption algorithm such as AES or KAES.Each phase of the proposed system is described in more details in the subsections.

1) Biometric Phase

A fingerprint represents a pattern of ridges and valleys on the person finger's tip. And also can be defined by the uniqueness of the local ridge characteristics and their relationships. Figure 2 shows these characteristics by marking minutiae points for the finger image. Minutiae points are these local ridge characteristics that occur either at a ridge ending or a ridge bifurcation [6].In low quality fingerprint images which contain noise and contrast deficiency usually results in pixel configurations similar to minutiae points. So, the automatic minutiae detection process becomes a difficult task [6].

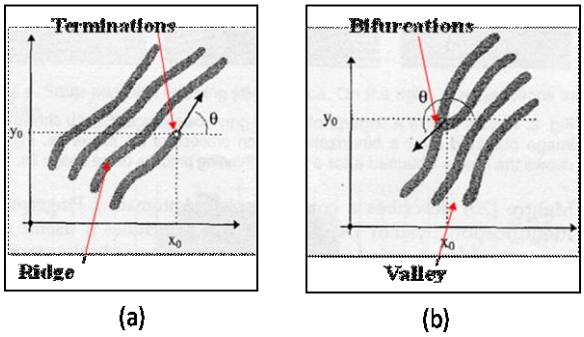


Fig.2. Fingerprint: (a) Termination minutia. (b) Bifurcation minutia.

The following steps are run in the biometric phase to form a biometric feature matrix using the input fingerprint image [6]:

1. Pre-Processing.
2. Minutiae extraction.
3. Post-Processing.

The first step called pre-processing and runs different tasks to enhance the input fingerprint image. The next step deals with the extraction of minutiae. In the third step false minutiae are deleted from the set of minutiae which obtained early. *Pre-Processing* step contains three different stages namely, image enhancement, image binarization and image segmentation. Each stag runs one or more different image processing methods as shown in figure 3. The input for this step is the fingerprint image and the output is an enhanced fingerprint image that is a suitable input for the following minutiae extraction step [6]. At *Minutiae extraction* step, the skeleton of the image is formed and the minutiae points are then extracted by the following method [6]:1. The binary image is thinned as a result of which a ridge is only one pixel wide.2. The minutiae points are thus those which have a pixel value of one (ridge ending) as their neighbor or more than two ones (ridge bifurcations) in their neighborhood

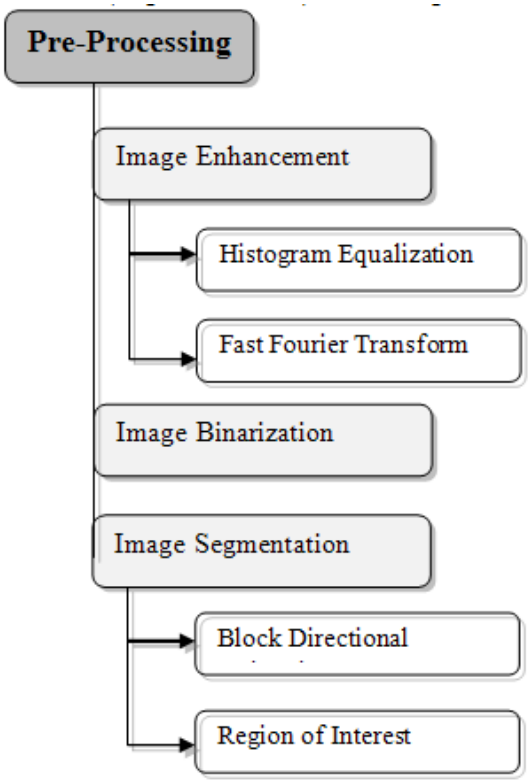


Fig.3. Pre-Processing Step

The output of this step is an Nx3 biometric feature matrix which contains x positions, y positions and orientations for N minutiae as:

X position	Y position	Orientation
112.0000	77.0000	3.1416
208.0000	34.0000	-1.3191
50.0000	54.0000	0.1326
177.0000	73.0000	2.4585
39.0000	55.0000	0.0286
157.0000	239.0000	-2.3816

Post-Processing step removes the false minutiae from the biometric feature matrix. These false minutiae may occur due to the presence of ridge breaks in the given image itself which could not be improved even after enhancement. Figure 4 shows different situations for false minutiae points.

The biometric phase outputs an Nx3 matrix which usually holds the true N minutiae points' information.

The biometric phase outputs an Nx3 matrix which usually holds the true N minutiae points' information.

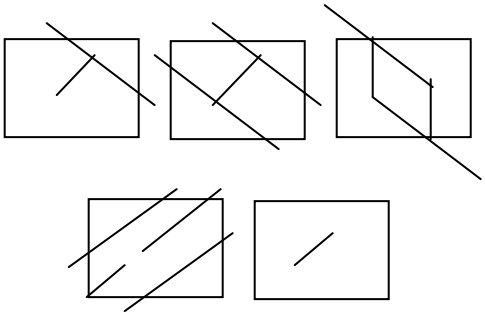


Fig.4. False Minutiae Points

The biometric phase outputs an Nx3 matrix which usually holds the true N minutiae points' information.

Key Generation Phase

In the proposed, MD5 is used to generate 128-bit encryption key from the generated biometric feature matrix .MD5 algorithm consist of 5 steps namely, Append Padding Bits, Append Length, Initialize MD Buffer, Process Message in 16-Word Blocks and Finalize the Output [7].The output generated key from MD5 is suitable for many encryption algorithms like AES and KASE.

Encryption Phase

KAES is a symmetric encryption technique that changes AES to be key dependent techniques. KAES is block cipher in which the block length and the key length are specified according to AES specification: 128, 192, or 256 bits and block length of 128 bits. In the proposed system, a key length of 128 bits is used. KAES involves the key in most of algorithm steps which increase the security of it rather than in AES. The main differences between AES and KAES can be summarized as in Table 1[5].Through the encryption phase, KAES has been applied to encrypt the plain-text using the generated key.

	AES	KAES
Round Function	last round is different	The same transformations in all rounds.
S-BOX	Fixed	S-Box is key dependent
Key Expansion	Fixed S-Box	Generate d S-Box
Round Transformation	Independent on the key	Dependent on the key
Shift Offset	Use Fixed from 0 to 3	Reliant on the key

Table 1. Differences between KAES and AES

III. Evaluation Criteria

For evaluating randomness of the proposed system, various tests were applied. To facilitate interpretation of the experimental results, a brief description is given, to make the analysis of these tests output understandable.

1) Entropy

Entropy describes the number of bits per byte. It is known as information density for a file. Extremely entropy output indicates that information is essentially random. Hence optimal compression is unlikely to reduce its size.

2) Optimal "Best" Compression

Reflects compressibility and is computed based on entropy encoding. For Example, if a file has 4.9 as Entropy, the optimal compression (OC) of the file would reduce its size by 38% as follow:

OC = 100 – (Entropy / 8)*100

OC = 100 – 62 = 38 %

3) Chi-square Distribution

The randomness of data can be tested using the chi-square test. The chi-square distribution is as an absolute number and a percentage which indicates how frequently a truly random sequence would exceed the value calculated.The percentage is interpreted as the degree to which the sequence tested is suspected of being non-random as:

- If the percentage is greater than 99% or less than 1%, the sequence is almost certainly Not Random.
- If the percentage is between 99% and 95% or between 1% and 5%, the sequence is Suspect.
- Percentages between 90% and 95% and 5% and 10% indicate the sequence is Almost Suspect.
- Otherwise the sequence is random.

4) Arithmetic Mean Value

Arithmetic mean value for a sequence of data is simply the result of summing the all the bytes and dividing by the sequences length. If the data are close to random, this should be about 127.5. If the mean departs from this value, the values are consistently high or low.

5) Monte Carlo value for PI

Each successive sequence of six bytes is used as 24-bit x and y coordinates within a square. If the distance of a randomly generated point is less than the radius of a circle inscribed within the square, the six byte sequence is considered a "hit". The percentage of hits can be used to calculate the value of PI. If the computed value approaches the correct value of PI, the sequence is close to random. Monte Carlo value for Pi is 3.143580574 (error 0.06 percent).

6) Serial Correlation Coefficient

The quantity measures the extent to which each byte in a sequence depends upon the previous byte. If the value (which can be positive or negative) close to zero, the

sequence is random (totally uncorrelated). Otherwise serial correlation coefficient will be greater than or equal 0.5.

IV. Experimental Results

To simulate the proposed biometric key dependent cryptosystem, a MATLAB script was implemented for biometric phase and for AES and KAES also a java program was implemented for key generation phases. The key's length (128 bit) was fixed for both AES and KAES algorithms. Fingerprint image (fingerprint3.tif) which captured by Cross Match Verifier 300 scanner at 500 dpi, is used to test our system. Figure 5 shows the inputs and outputs of the biometric phase program. The minutiae points are saved as a biometric feature matrix.

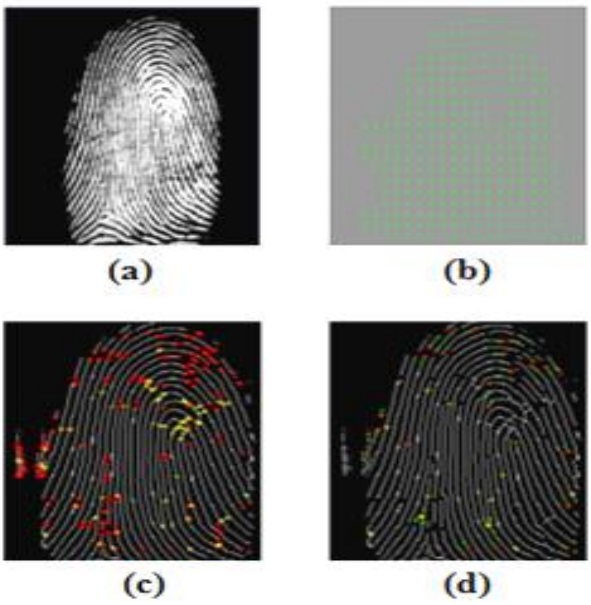


Fig.5. Biometric Phase: (a) Input Image (b) Preprocessing output (c) Extraction output (d) Post processing output

The generated 128 key from the biometric features matrix through the key generation phase is:
0X9e 0Xf8 0X68 0X62
0X53 0Xdf 0X3c 0Xc5
0Xf2 0X66 00Xef 0Xde
0Xb6 0Xc7 0Xbe 0X3a

For the testing another plain key is used:
0Xb1 0Xc2 0Xf3 0X84
0X75 0Xa6 0Xd7 0X08
0X19 0X13 0X11 0X42
0X53 0X20 0X15 0X16

Table 2 lists six files with their format and sizes. These files are used in encryption and decryption steps. Using ENT [8] results is obtained by carrying out the evaluation criteria discussed in section 3.

	Name	**Size** (Bytes)	**Name**
Text	secret_t1.txt	2815	T1
	secret_t2.txt	4007	T2
Audio	secret_w1.wav	14077	W1
	secret_w2.wav	35191	W2
Image	secret_im1.tif	95162	Im1
	secret_im2.tif	1001648	Im2

Table 2. Files Names

Table 3 illustrates the occupation of bits within a byte, it could be noticed that KAES and biometric KAES utilizes almost every bit in a byte, introducing extremely dense files.Figure 6 depicts the optimum "best compression" that can be achieved. The results are the same for some of the experimented files.

The Chi-Square distribution for the experimental file is shown in figure 4. Also Chi-Square distribution gives a good indication of the proposed system randomnTable 4 shows the computed arithmetic mean for both AES and KAES are close to the arithmetic mean value = 127.5.Table 5 Shows the computed Monte-Carlo values of piwhich are close to the value of PI= 3.1416.Figure 8 indicates that relation between successive bytes is very small as Serial Correlation Coefficients

.

	T1	T2	W1	W2	Im1	Im2
Original	4.711491	5.166631	6.215379	5.963367	5.910927	7.562718
KAES_PlainKey	7.911388	7.952993	7.98213	7.986803	7.996512	7.999835
KAES_Fingerprint key	7.930583	7.947021	7.983541	7.985064	7.996409	7.99978
AES_PlainKey	7.926212	7.957024	7.984306	7.983461	7.997254	7.999767
AES_Fingerprint key	7.918981	7.955877	7.985193	7.985132	7.996956	7.999767

Table 3. Entropy Values

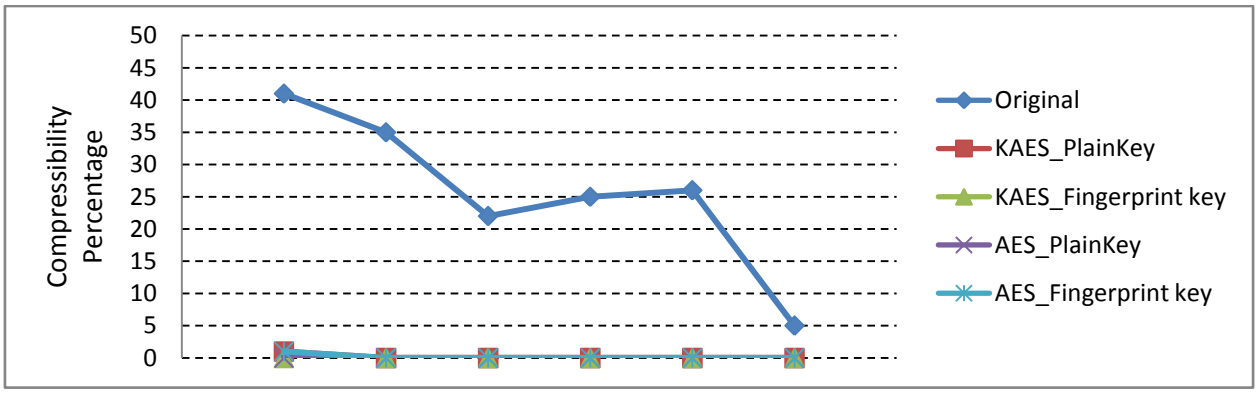


Fig.6. The Optimum Compression

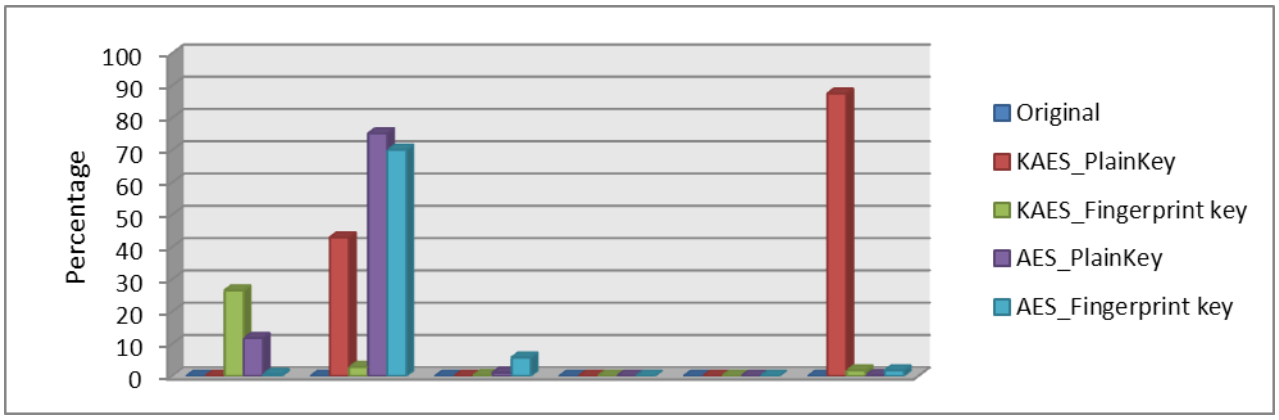


Fig.7. The Chi-Square distribution

	T1	T2	W1	W2	Im1	Im2
Original	85.5844	133.5031	127.0035	121.4201	65.9731	93.1652
KAES_PlainKey	127.4348	124.9963	127.1517	127.1777	127.0659	127.5825
KAES_Fingerprint key	127.0274	128.4168	127.5046	128.2845	127.469	127.5014
AES_PlainKey	130.323	127.2298	127.7911	128.4338	127.58	127.6038
AES_Fingerprint key	128.6254	128.1158	127.7079	127.3429	127.4235	127.6049

Table 4. The Computed Arithmetic Mean

	T1	T2	W1	W2	Im1	Im2
Original	4	2.824587706	3.884057971	3.957033248	3.384867591	3.865725017
KAES_PlainKey	2.950959488	3.166416792	3.15942029	3.156351236	3.15889029	3.136844754
KAES_Fingerprint key	3.138592751	3.274362819	3.15771526	3.125660699	3.161916772	3.14951989
AES_PlainKey	3.068376068	3.076461769	3.109974425	3.144075021	3.157124842	3.136293661
AES_Fingerprint key	3.025641026	3.148425787	3.15771526	3.118158568	3.145775536	3.136293661

Table 5. The Monte-Carlo Values

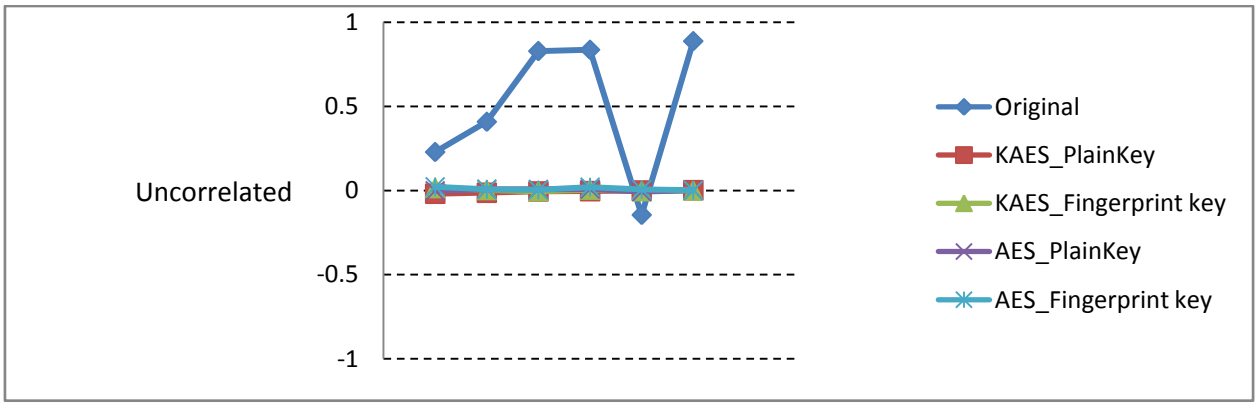


Fig.8. The Serial Correlation Coefficient

V. Conclusion

This paper presents a biometric key dependent cryptosystem by replacing the plain key with fingerprint feature data. KAES is improving the security of the proposed system by employing the key to be the main parameter of the encryption algorithm. Experiments analysis for biometric and key generation phases will be reported in the near future to insure the reliability and the security of the proposed system.

VI. References

1) International Computer Security Association., & Nichols, R. K. (1999). ICSA guide to cryptography (chapter 22). New York: McGraw Hill.
2) Stoianov, A., Information and Privacy Commissioner /Ontario., & Cavoukian, A. (2007). Biometric encryption: A positive-sum technology that achieves strong authentication, security and privacy. Toronto, Ont: Information and Privacy Commissioner, Ontario.
3) Kresimir & Mislav(2004, June). A survey of biometric recognition methods. presented at 46th International Symposium Electronics in Marine, ELMAR-2004
4) Li, W., Zhan, C., & Zheng, G. (January 01, 2006). Cryptographic Key Generation from Biometric Data Using Lattice Mapping. Proceedings, 513-516.
5) Uludag, U. (2006). Secure biometric systems.
6) Nimitha Chama(2003). Fingerprint Image Enhancement and Minutiae Extraction. University of Clemson.
7) Network Working Group, R. Rivest & RSA Data Security. Retrieved July 25, 2010 from Internet FAQ Archives Web site: http://www.faqs.org/rfcs/rfc1321.html
8) ENT. Retrieved July 20, 2010 from A Pseudorandom Number Sequence Test Program Web site: http://www.fourmilab.ch/random

Economical Task Scheduling Algorithm for Grid Computing Systems

Amit Agarwal, Padam Kumar

GJCST Classification
D.4.1, F.1.2

***Abstract*- Task duplication is an effective scheduling technique for reducing the response time of workflow applications in dynamic grid computing systems. Task duplication based scheduling algorithms generate shorter schedules without sacrificing efficiency but leave the computing resources over consumed due to the heavily duplications. In this paper, we try to minimize the duplications of tasks from the schedule obtained using an effective duplication based scheduling heuristic without affecting the overall schedule length (makespan) of grid application. Here, we suggested an economical duplication based intelligent scheduling heuristic called economical duplication scheduling in grid (EDS-G). The simulation results show that EDS-G algorithm generates better schedule with lesser number of duplications and remarkably less resource consumption as compared with HLD, LDBS in the simulated heterogeneous grid computing environments**

Keywords- scheduling, grid computing, duplication based heuristic, DAG, workflow applications.

I. Introduction

In recent years, grid computing has obtained a lot of attentions from engineers and scientists for executing high performance parallel and distributed applications due to major advancements in wide-area network technologies and low cost of powerful computing and high-speed network resources. In general, a parallel and distributed application can be represented by a weighed directed acyclic task graph (DAG) as shown in figure 1. In DAG, the nodes represent application tasks and the edges represent inter-task data dependencies. The algorithm for finding an optimal schedule for the multiprocessor scheduling problem is NP-complete [1, 2, 3]. In literature, task scheduling heuristics for DAG applications have been classified as list scheduling [4, 5, 6] cluster-based scheduling [7, 8] and duplication-based scheduling [9, 10, 11, 12, 13, 14, 15, 16, 17]. In list-based task scheduling, tasks are ordered in non-increasing order of their priorities (or ranks) and scheduled on the resource which minimizes the objective function such as schedule length. Clustering is an efficient way to reduce communication delay in DAGs by grouping heavily communicating tasks to same labeled cluster and then assigning tasks in a cluster to the same resource. In duplication-based scheduling, parents of current selected task can be duplicated into idle time slots between two already scheduled tasks in order to reduce the task finish/ start time. Duplication-based scheduling is very effective in distributed computing system but leave the computing resources over consumed due to the heavily duplications.

About-Department of Electronics & Computer Engineering Indian Institute of Technology, Roorkee (India) {aamitdec, padamfec} @iitr.ernet.in

Duplication heuristics are more effective for fine grain task graphs and for networks with high communication latencies. The term CCR refers to the ratio of average communication cost to average computation cost on a given system. A high CCR indicates the communication intensive nature of a problem, whereas, low CCR represents the computation intensive problem. Duplication plays its role more effectively at higher CCRs, as the formation of large sized scheduling holes increases with higher communication costs, which can be exploited to accommodate fine grain tasks conveniently [10]. In heterogeneous distributed computing system, heterogeneity of computational resources and communication mechanisms poses some major obstacles to achieve high parallel efficiency. Performance of the scheduling algorithms tends to degrade in the presence of heterogeneity. This degradation becomes more pronounced with an increase in heterogeneity and at higher CCRs which results in inappropriate task/ processor selection. In this case, duplication is very graceful to overcome these „stresses' and „strains' of heterogeneity by duplicating the crucial tasks and thereby improving the finish time on processing resource, but it increases scheduling cost due the duplicated tasks overhead. In [16], Savina et al. suggested the heterogeneous limited duplication (HLD) that adapts the SD algorithm [10] heterogeneous environment and then assessed the usefulness of limited duplication approach in dealing with the stresses of heterogeneity in a system. In [17], Dogan et al. proposed a level sorting algorithm (LDBS) to arrange the tasks in DAG into various precedence levels. The tasks belonging to the same level have no data dependencies can be executed concurrently. In LDBS, tasks are scheduled level by level starting from the top. In current economic market models [18, 19], economic cost (cost of executing a workflow on grid) has been considered as an important scheduling criterion to employ the user-centric policies, since different resources, belonging to different organizations, may have different polices of charging. Hence, the economic cost of resource consumption for scheduling the grid applications becomes an important performance metrics for analyzing the scheduling algorithms.In EDS-G algorithm, we analyze the impact of the duplicated tasks over makespan and try to optimize the schedule generated with duplication by eliminating tasks (duplicated and unproductive) as much as possible without affecting the makespan. A task may become unproductive after being duplicated if its immediate successor tasks can gather output data from its duplicated parent task not later than the unproductive task. These algorithms can prove that they are very useful in the distributed grids to reduce the scheduling cost of an

application and improving the performance of the grid system.The remainder of this paper is organized as follows. Section II defines the task scheduling problem. Section III presents the proposed task scheduling algorithms. Section IV shows the simulation results. Then, in section V, we describe our conclusion of current research work.

II. TASK SCHEDULING PROBLEM

A task scheduling model for grid computing system includes an application of dependent tasks (DAG); a target grid computing system of arbitrary connected multiple computing resources and an objective function.

1) *Grid Resource Model*

A grid computing system can be represented by $G=(R,Q)$ where R is the set of *m* arbitrary connected computing resources $(r_1, r_2, ...r_m)$ forming a grid and Q is the set of communication channels connecting the grid resources. In grid computing system, task execution cost on different grid resources may be different due to the processor heterogeneity (different processing rates of grid resources) and similarly, the data transfer rates (bandwidths) between different pair of processing resources may be different due to network heterogeneity. In this model, it is also assumed that each processing resource has co-processor to deal with communications, which allows computation and communication to overlap each other. Additionally, task executions are assumed to be non-preemptive and communication overhead between two tasks scheduled on the same resource is considered as zero. After completing execution of a task, the associated grid node sends output data to all of its child tasks in parallel.The main objective of this paper is to minimize duplications after duplicating tasks over grid resources selectively and minimize the overall schedule cost (Resource Consumption). A resource consumption can be defined as the fraction of time duration (between after being allotted to an application and released for another application) a resource is actually executing some tasks of application in grid.

Task Node	Computation costs on different grid nodes				Mean cost
	r_1	r_2	r_3	r_4	$\bar{\tau}_i$
n_1	1	1	2	1	1.25
n_2	3	2	4	2	2.75
n_3	5	6	3	4	4.5
n_4	2	4	4	2	3.0
n_5	4	8	7	8	6.75
n_6	3	3	1	2	2.25
n_7	5	5	5	5	5.0
n_8	1	2	2	2	1.75

Table 1. Computation cost matrix [τ_{ij}] for DAG in fig. 1.

2) *Grid Application Model*

A grid application may be represented by a weighted directed acyclic graph (see figure 1) or DAG, $D=(N,E,T,C)$ where N is a set of n computation task nodes, T is a $n \times m$ computation cost matrix (see table 1) and the value of $\tau_{ij} \in T$ is theexpected time to execute task n_i on grid resource r_j *for* $1 \le i \le n$ and $1 \le j \le m$, E is a set of communication edges that shows precedence constraints among the tasks and C is a $n \times n$ communication cost matrix and the value of $c_{ij} \in C$ is the expected time to communicate data from task n_i to task n_j for $1 \le i \ne j \le n$. The mean computation cost $\bar{\tau}_i$ of task n_i and mean communication cost $\bar{c}_{ij}$ between task n_i and task n_j can be calculated as

$$\bar{\tau}_i = \frac{\sum_{j=1}^{m} \tau_{ij}}{m} \qquad \forall 1 \le i \le n \qquad (1)$$

$$\bar{c}_{ij} = \frac{c_{ij}}{mean\ data\ transfer\ rate\ over\ all\ links\ in\ grid} \qquad \forall\, 1 \le i \ne j \le n \qquad (2)$$

A task node without any parent node is called entry task and a task node without any child node is called exit task. If there are two or more entry (exit) tasks, they may be connected to a zero-cost pseudo entry (exit) task with zero-cost edges which will not affect the schedule. Since the intra-processor bus speed is much higher than the inter-processor network speed, the communication cost between two tasks scheduled on the same processing node is considered as zero [20].

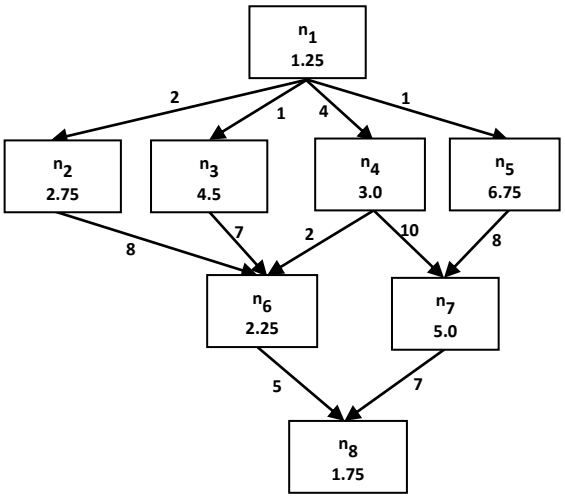


Fig. 1. A Simple DAG with Precedence Constraints.

III. ECONOMICAL DUPLICATION BASED SCHEDULING

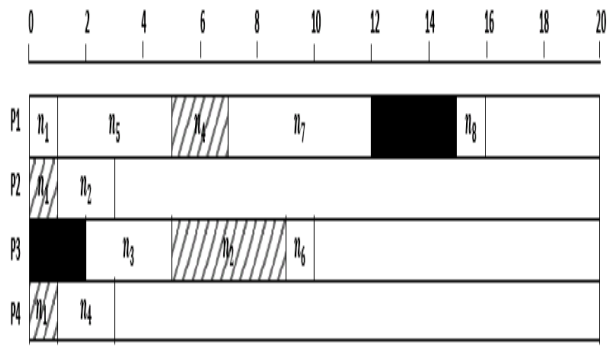
This section presents the economical duplication based scheduling algorithm (EDS-G) in grid computing environment inspired from our earlier work in [21]. This algorithm consists of two mechanisms, first is a lower-bound complexity mechanism for scheduling based on insertion based task duplication and second is modifying schedule after removing some duplicated and unproductive tasks in the schedule without affecting the makespan.In this section, a lower-bound complexity algorithm (EDS-G) for grid computing system has been presented. The pseudo code of the algorithm is shown in figure 3. A priority-based task sequence is generated by ordering the tasks in non-increasing order of their b-level (computation and communication cost along the longest directed path from the concerned task to the exit task in DAG) that can be calculated recursively using mean cost parameter as:

$$b_i = \bar{\tau}_i + \max\{b_j + \bar{c}_{ij}\} \quad \forall n_j \in succ(n_i) \qquad (3)$$

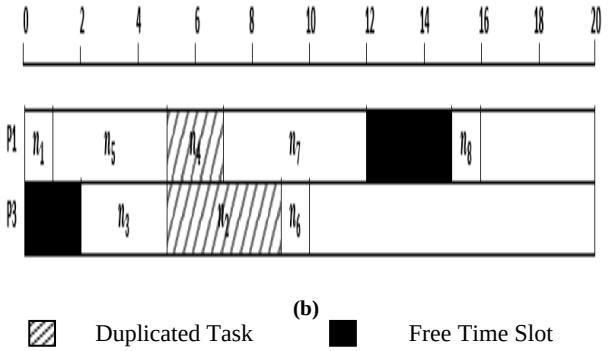
The term $succ(n_i)$ refers to the set of immediate child nodes of task n_i in the DAG. Now, the first unscheduled task in the task sequence is selected and scheduled on a grid resource that can finish its execution at the earliest using duplication (task replication) approach. This algorithm uses insertion based scheduling policy which considers the possible insertion of a task or duplicated task in an earliest idle time slot between two already scheduled tasks on the grid resource. A task on the grid resource can start execution only after the data arrived from all of its immediate predecessors. The parent of task n_i whose data arrives last of all is termed as the most important immediate parent (MIIP). Data arrival time for n_i on r_k is given by:

$$DAT(n_i, r_k) = \max_{n_j \in pred(n_i)} \{\min\{F_{jk}, F'_{jk} + c_{ji}\}\} \qquad (4)$$

The term $pred(n_i)$ refers to the set of immediate parent nodes of task n_i in the DAG.



(a)



(b)

Duplicated Task Free Time Slot

Fig. 2. Gantt Charts for the schedule generated by (a) HLD Algorithm (Duplications = 4, Resource Used = 4) (b) EDS-G Algorithm (Duplications=2, Resource Used = 2) for an application DAG shown in fig. 1.

Due to the non-availability of data earlier, owing to precedence constraints or communication delay, a grid resource may remain idle leading to the formation of scheduling holes. These scheduling holes may be exploited to duplicate tasks to minimize data arrival time. The start time S_{ik} of task n_i on grid resource r_k is limited by the data arrival from its MIIP (say M_i) and availability of a suitable scheduling hole. If suitable slot is not available then task n_i can start after the completion of last scheduled task on grid resource r_k, i.e. the ready time (r_k^R) of resource r_k. The start time of task n_i on resource r_k is given by:

$$S_{ik} = \max\{DAT(M_i, r_k), \min\{r_k^R, G_r^S\}\} \qquad (5)$$

where G_r^S is the start time of first suitable free time slot G_r to accommodate task n_i on the resource r_k, if exist. The finish time F_{ik} is calculated as:

$$F_{ik} = S_{ik} + \tau_{ik} \qquad (6)$$

The finish time is calculated for all the available grid resources and task n_i is scheduled on the resource that gives earliest finish time. After scheduling the all tasks, makespan is calculated as:

$$makespan = \max\{F_{ik}\} \quad \forall 1 \le i \le n; 1 \le k \le m \qquad (7)$$

EDS-G Algorithm

Begin

1: Construct a priority based task sequence ξ ;
2:do {

3: Select the first unscheduled task n_i in the task sequence ξ

4:for (all r_k in processor list R) {

5: Sort the list of immediate parents of n_i in non-increasing order of data arrival time;
6:for all immediate parents, select the first immediate parent

n_j from the list at Step 5 {

7:if duplication of n_j can reduce finish time F_{ik} of n_i on r_k

8: Duplicate n_j ;
9: }

10: Compute earliest finish time F_{ik} of n_i on r_k using eq.(6);
11: }

12: Find the minimum earliest finish time of n_i ;

13: Assign n_i on resource r_k with minimum F_{ik} in schedule **S**;

14: }**while** (there are unscheduled tasks in the task sequence ξ);
15: Maintain a list **A** of origin tasks which have been duplicated later with their successor links to other tasks and list **B** of duplicated tasks in non-increasing order of their earliest start time;

16:for (each duplicated task n_i in list **B**) {
17:if (no change in makespan of schedule **S** after removing

duplicated task n_i)

18: Remove this duplicated task n_i from the schedule **S** and update list **A**;
19: }

20:for (each task n_i in list **A**) {

21:if task n_i has no dependent task in schedule **S** due to its

duplication later on different processors; remove this task n_i from the schedule **S**
22: }

Fig. 3. Pseudo Code for EDS-G Algorithm

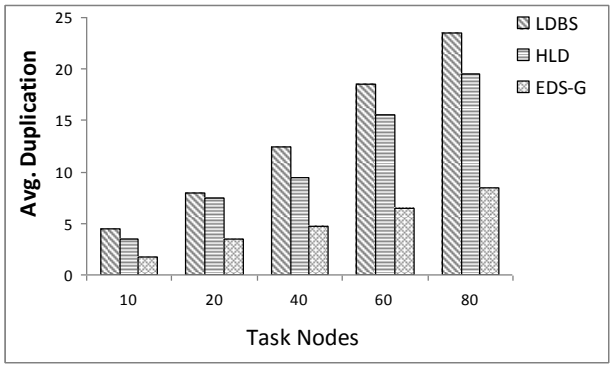
Further, we maintain a list A of origin tasks (which have been duplicated later in the schedule) with their links to dependent tasks and list B of duplicated tasks in non-increasing order of earliest start time. The above schedule is modified only if the removal of the duplicated task from list B does not affect the makespan. Similarly, all the unproductive tasks in the list A, that are not responsible to provide any output to immediate successor tasks due to theirduplications, are removed from the schedule. This modified schedule contains lesser number of duplications and remarkably less resource consumption as compared with HLD, LDBS for heterogeneous grid computing systems. Gantt charts for the schedule generated by HLD algorithm and EDS-G Algorithm is shown in fig. 2. Here in HLD schedule, task n_2 on resource r_2 and task n_4 on resource r_4 are unproductive after being duplicated on resource r_2 and r_1 respectively. Hence in EDS-G schedule, tasks n_1 (which was scheduled on r_2 for n_2), n_2 from resource r_2 and tasks n_1 (which was scheduled on r_4 for n_4), n_4 from resource r_4 have been removed. It shows that EDS-G uses less duplications and lesser number of resources as compared to HLD for the same makespan.

IV. Simulated Results and Analysis

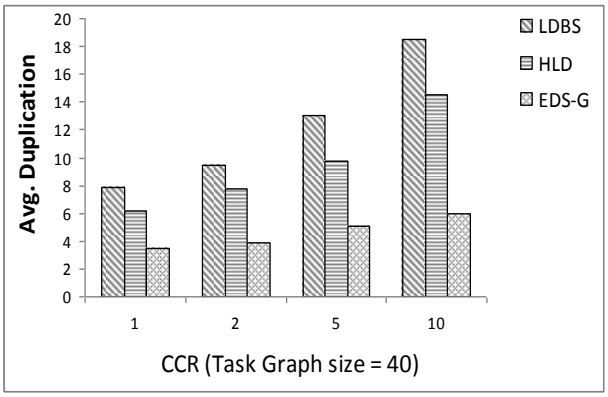
The experimental results for random task graphs are presented in fig. 6 and 7 for the clique topology for different task graph sizes and CCRs in the simulated grid environments. The performance of EDS-G algorithm is analyzed with respect to various graph characteristics (task sizes, CCRs). The simulated set of experiments compare the performance of the grid system in terms of average number of duplications and resource consumption with respect to various graph sizes and CCRs for heterogeneous grid computing systems (see fig. 6). Each result is obtained with respect to CCR is an average of 25 graphs (over 5 sizes and 5 average parallelisms), and with respect to graph size is an average of 20 graphs (over 4 CCRs and five average parallelisms). In these experiments, the EDS-G algorithm outperforms the HLD, LDBS for different DAG sizes and CCRs with respect to avg. duplications and resource consumption.

V. Conclusions

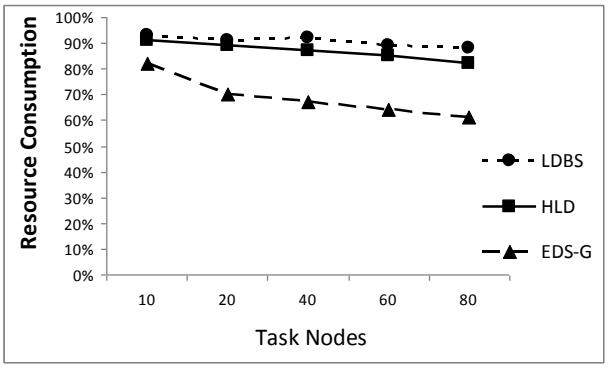
A duplication based strategy has been found very momentous for homogeneous and heterogeneous computing systems to improve the performance of the system. Currently more research work is focusing over heterogeneous computing system such as grids which consists of abundant resources over wide area networks with different capabilities. Scheduling a task graph with precedence constraints in grids is an issue due to the higher communication latencies. Duplication improves performance and reliability of such systems by duplicating critical tasks with higher communication costs.Our approaches optimize schedule by reducing duplications as much as possible without affecting the makespan and improve the system performance so that application execution cost and duplication overhead can be reduced. Performance comparison with best known duplications algorithms LDBS and HLD for grid computing system shows that EDS-G algorithm generates comparable schedules with remarkably less duplications and less resource consumption.



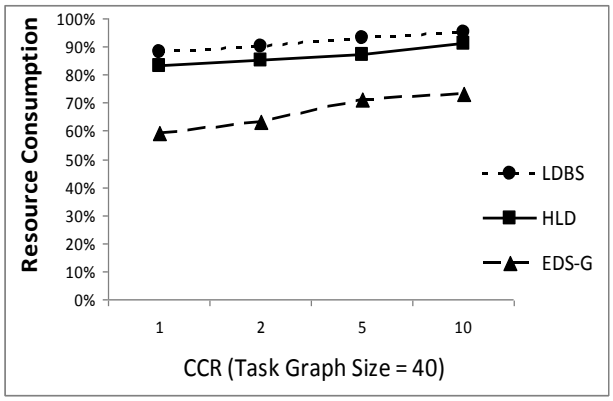
(a)



(b)



(c)



(d)

Fig 6. (a to d) Performance comparison of EDS-G algorithm on random DAGs in computational Grids.

VI. References

1) Y.-K. Kwok and I. Ahmed, Benchmarking the Task Graph Scheduling Algorithms, IPPS/SPDP, pp. 531-537, 1998.
2) J. Liou and M. Palis, A Comparison of General Approaches to Multiprocessor Scheduling, Proc. of the 11th Int'l Parallel Processing Symp., pp. 152-156, 1997.
3) A. A. Khan, C. McCreary and M. S. Jones, A Comparison of Multiprocessor Scheduling Heuristics, ICPP, pp. 243-250, 1994.
4) Olivier B, Vincent B and Yves R, The Iso-level Scheduling Heuristic for Heterogeneous Processors, Proceedings of 10th Euromicro workshop on parallel, distributed and network-based processing, pp. 335-342, 2002.
5) H. Topcuoglu, S. Hariri and M.Y. Wu, Performance-Effective and Low-Complexity Task Scheduling for Heterogeneous Computing, IEEE Trans. on Parallel and Distrib. Syst., 13(3): pp. 260 - 274, March 2002.
6) T. Hagras and J. Janecek, A High Performance, Low Complexity Algorithm for Compile-Time Job Scheduling in Homogeneous Computing Environments, Proc. Int'l Conf. on Parallel Processing Workshops, pp 149-155, Oct 2003.
7) A. Gerasoulis and T. Yang, A Comparison of Clustering Heuristics for Scheduling Directed Acyclic Graphs on Multiprocessors, Journal of Parallel and Distributed Computing, 16(4): pp. 276-291, 1992.
8) J. Liou and M. A. Palis, An Efficient Task Clustering Heuristic for Scheduling DAGs on Multiprocessors, Proc. of Workshop on Resource Management, Symposium of Parallel and Distributed Processing, pp. 152-156, Oct 1996.
9) Kuan-Chou Lai and Chao-Tung Yang, A Dominant Predecessor Duplication Scheduling Algorithm for Heterogeneous Systems, Journal of Supercomputing, 44(2): pp. 126-145, 2008
10) Savina Bansal, Padam Kumar and Kuldip Singh, An Improved Duplication Strategy for Scheduling Precedence Constrained Graphs in Multiprocessor Systems, IEEE Trans. on Parallel and Distributed Systems, 14 (6): pp. 533-544, 2003.
11) I. Ahmed and Y.-K. Kwok, On Exploiting Task Duplication in Parallel Program Scheduling, IEEE Trans. on Parallel and Distributed Systems, 9(9): pp. 872-892, Sept 1998.
12) G.-L. Park, B. Shirazi and J. Marquis, DFRN: A New Approach for Duplication Based Scheduling for Distributed Memory Multiprocessor Systems, Proc. of the 11th Int'l Parallel Processing Symp., pp. 157-166, Apr 1997.
13) Y.-C. Chung and S. Ranka, Application and Performance Analysis of a Compile-Time Optimization Approach for List Scheduling Algorithms on Distributed Memory

Multiprocessors, Proc. on Supercomputing, pp. 512-521, Nov 1992.
14) B. Kruatrachue and T.G. Lewis, Grain Size Determination for Parallel Processing, IEEE Software, 5(1): pp. 23-32, Jan 1988.
15) C.H. Papadimitriou and M. Yannakakis, Towards an Architecture-Independent Analysis of Parallel Algorithms, SIAM J. Computing, 19(2): pp. 322-328, Apr 1990.
16) Savina Bansal, Padam Kumar and Kuldip Singh, Dealing with Heterogeneity Through Limited Duplication for Scheduling Precedence Constrained Task Graphs, Journal of Parallel and Distributed Computing, 65(4): 479-491, Apr 2005.
17) A. Dogan and F. Ozguner, LDBS: A Duplication Based Scheduling Algorithm for Heterogeneous Computing Systems, Proceedings of the Int'l Conf. on Parallel Processing, pp. 352-359, Aug 2002.
18) C. Ernemann, V. Hamscher, and R. Yahyapour, Economic scheduling in grid computing, Proceedings of the 8th Workshop on Job Scheduling Strategies for Parallel Processing, Vol. 2537 of Lecture Notes in Computer Science, Springer Verlag, pp.128–152, 2002.
19) J. Yu, R. Buyya, and C. K. Tham, Cost-based scheduling of scientific workflow application on utility grids, Proceedings of the First IEEE International Conference on e-Science and Grid Computing (e-Science 2005), pp.140–147, IEEE Computer Society, 2005.
20) Mohammad I. Daoud and Nawwaf N. Kharma, Efficient Compile-Time Task Scheduling for Heterogeneous Distributed Computing Systems, Proceedings of the 12th IEEE International Conference on Parallel and Distributed Systems, pp. 11-22, 2006.
21) A. Agarwal and P. Kumar, Economical duplication based task scheduling for heterogeneous and homogeneous computing systems, IEEE International Advance Computing Conference (IACC 2009), pp.87–93, 2009.

Student Relationship in Higher Education Using Data Mining Techniques

GJCST Classification H.2.8, F.4.2

Boumedyen Shannaq[1], Yusupov Rafael[2], V. Alexandro[3]

Abstract- **The aim of research paper is to improve the current trends in the higher education systems to understand from the outside which factors might create loyal students. The necessity of having loyal students motivates higher education systems to know them well, one way to do this is by using valid management and processing of the students database. Data mining methods represent a valid approach for the extraction of precious information from existing students to manage relations with future students. This may indicate at an early stage which type of students will potentially be enrolled and what areas to concentrate upon in higher education systems for support. For this purpose the data mining framework is used for mining related to academic data from enrolled students. The rule generation process is based on the decision tree as a classification method. The generated rules are studied and evaluated using different evaluation methods and the main attributes that may affect the student's loyalty have been highlighted. Software that facilitates the use of the generated rules is built using VB.net programming language which allows the higher education systems to predict thestudent's loyalty (numbers of enrolled students) so that they can manage and prepare necessary resources for the new enrolled students.**

Keyword-Data mining ,decision tree , Exploratory data analysis, Adaptive System .

I. Introduction

Nowadays there is an evolution of educational systems and there is a great importance of the educational field. Modern educational organizations start developing and enhancing the educational system increasing their capability to help the decision makers obtain the right knowledge, and to make the best decisions by using the new techniques such as data mining methods [1]. Subsequently, a suitable knowledge needs to be extracted from the existing data. Data mining is the process of extracting useful knowledge and information including: patterns, associations, changes, anomalies and significant structures from a great deal of data stored in databases, data warehouses, or other information repositories. The data mining expediency is delivered through a series of functionalities such as outlier analysis, evolution analysis, association analysis, classification, clustering and prediction. Data mining is an integral part of Knowledge Discovery in Database (KDD) [2][3]. Student enrollment process in any higher education system is of great concern of the higher education managements. Severalfactors may affect the student enrollment process in a particular Institute. One of the biggest challenges that higher education faces today is predicting the paths of loyal students (enrolled students). Institutions would like to know, for example, which students and how many will enroll in particular institute.This research paper is an attempt to use the data mining processes, particularly predictive classification to enhance the quality of the higher educational system to increase numbers of loyal students (enrollment students) to evaluate student data to study the main attributes that may affect the student enrollment factors to plan for institutes resources (Instructors, classes, labs, etc.)from knowing how many students will be enrolled and to make a big effort to concentrate on all factors that play main role in motivating the new student to enrolls in a particular institute.

About[1]. Information Systems Department University of Nizwa , Sultanate of OmanEmail: boumedyen@unizwa.edu.om Email: boumedians@yahoo.com
About[2]. Director of Saint Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences Saint Petersburg, Russia Email: yusupov@iias.spb.su
About[3]- Academician of Academy of Natural Sciences, Russia Doc.Nat.Sci., St.Petersburg Institute Doc.Nat.Sci., St.Petersburg Institute E-mail: alexandr@iias.spb.su

II. Related Work

Higher education enrollment and admission departments are increasingly being asked to do more work in these aspects. Additionally, as described in [4] they recognize that each institution has different student relationship management, admissions and enrollment goals. In strengthening student relationship management for enterprise students projects, the project aims to improve the overall quality of the "entrepreneurial" student experience by mapping existing and proposing new service designs, this can lead to an increase in the efficiency and effectiveness of University or College's system supporting the management of relationships with students as well as in the interface between operational processes and the platforms and technical solutions used for it. As described in Student Relationship Management (SRM) article they introduce the topic of Student Relationship Management (SRM) in Germany. The concept has been derived from the idea of a Customer Relationship Management (CRM), which has already been successfully implemented in many enterprises [4]. The German university information system HIS ascertained that 65% of the first-year students decided to choose their university due to their place of residence. Good equipment of the university is an important criterion for 58% of high school graduates, as well as the reputation of the university or college (52%). Nevertheless, for 90% of the high school graduates it is above all important that the courses offered correspond to their specialized interests [5]. Furthermore, the use of information offered by university rankings becomes more and more frequent. However, not only German but also foreign students were used to select German university on the basis of the results of university

rankings which are available for the people. Therefore, it is not amazing that faculties with good ranking results register more students in the following term in relation to the previous years. However, neither the university management nor education administrators are aware of this competition trend because they associate competition rather with areas of research and professors than with students. Good universities, though, are not only characterized by outstanding researchers but by excellent students as well [6]. The research has [7] proposed a model to represent how data mining is used in higher educational system to improve the efficiency and effectiveness of the traditional processes in this model a guideline was presented for higher educational system to improve their decision-making processes. In [8] a complete system implemented for parsing such repositories into a SQL database and for extracting from the database and repositories, various statistical measures of the code and version histories. The research by [9] is to use Rough Set theory as classification approach to analyze student data. They build a higher educational system to improve their decision-making procedures through data mining. Other papers discuss different Al technologies and compare them with genetic algorithm based induction of decision trees and discuss why the approach has a potential for developing into an alert tool. Many other related works can be found in [12, 13, 14, 15, 16, and 17].

III. Data pre-processing

In the entire data mining process, the data cleaning process is utilized in order to eliminate irrelevant items. The discovery of patterns will be only useful if the data represented in files offer a real representation of the enrolment process and the actions or decisions taken by the past student [16]. Here we study the enrollment factors and decisions taken by students for enrollment in a particular institute. Subsequently if the student is enrolled in a particular institute then he is considered as a loyal student. Initially, the institute provided us a database equal to 19012. After filtering process of the data 2106 records were left. The data supplied is from a particular institute, we cannot mention its name, but it is from the Arabic region. The main objective of our research is to discover patterns that will be used to predict a loyal or not loyal student previously to his enrollment or not enrollment to a particular institute. By taking information from other students with similar information, in this sense, we can know the role of each attribute and the implicit relations among them.

IV. Exploratory data analysis

To achieve the goals, we analyze the behavior over time of students who first register and pay fees. We did that to eliminate possible effects in case of changing the structure of the institute. We first considered all students that have entered into the students database between 2003 and 2007, the total number is large, it is time consuming to analyze the whole data set, so we took a sample and analyzed that. We selected the same number of students from each time slot over the whole entry period; the sample contains a total of 2069 students. Generally it is not necessary to sample the data; the main reason for doing it here is the low quality of available data. After a long process of data management we obtained the variables, the features (variables) listed in the data set are the following:

Age: three classes for age between 18 – 28 (young) given number one in data set, for age between 29 -39 (middle) given number two and for age above 39 given number 3 for representation data in the data set.

Ch: number of children's

Sx : define female and male, codes of 0 and 1 for female and male.

Rel : define religion, codes of 0 and 1 for Muslim and other .Pob: place of birth, codes of 0 and 1 for original (residence) and foreign.

Nationality: codes of 0 and 1 for original(residence) and foreign.

SAvg: grade of secondary school.

Stype: type of study, codes of 0 and 1 for literal and scientific.

PGS: place of graduated secondary school, codes of 0 and 1 for original(residence) and foreign.

Numos: Numbers of different specialization selected by the new student.

RPf: student information entered in student database and the student pays for some registered courses.

Enroll: the student enrolled has codes 1 and for student who did not enroll with 0.

RRe: reason of enrollment in a particular institute codes of 1,2,3,4,5,6,7,8,9 for Public institute, much specializations ,accept low average after secondary school, much (prof, dr.), private institute , near place, much graduated students from this institute gets a job and offered grants .

Loyalty : codes 1 and 0 for loyal and not loyal.

We assigned descriptive value labels for each value of a variable. This process is particularly useful if your data file uses numeric codes to represent non-numeric categories (for example, codes of 1 and 0 for male and female).Value labels are saved with the data file. You do not need to redefine value labels each time you open a data file, the value labels process illustrated in figure4.1 and figure4.2.

age	Ch	sx	Rel	pob	nationality	SAvg	Stype	PGS	numos	RPf	Enrol	RRE	loyality
3	1	0	1	1	1	77.0	0	1	4	1	1	7	1
1	0	1	0	0	0	50.0	1	0	4	0	0	6	0
2	1	0	0	0	0	59.8	1	0	4	4	1	7	1
3	3	1	1	1	1	65.9	1	1	4	4	1	7	1
1	0	0	1	0	0	71.4	0	1	4	3	1	7	1
1	0	1	1	0	0	68.3	1	1	4	4	1	7	1
1	0	0	1	0	0	58.0	1	1	4	4	1	7	1

Figuer4.1: Data set Sample in label view

INTRODUCTION

age	Ch	sx	Rel	pob	nationality	SAvg	Stype	PGS	numos	RPf	Enrol	RRE	loyality
3	1	0	1	1	1	77.0	0	1	4	1	1	7	1
1	0	1	0	0	0	50.0	1	0	4	0	0	6	0
2	1	0	0	0	0	59.8	1	0	4	4	1	7	1
3	3	1	1	1	1	65.9	1	1	4	4	1	7	1
1	0	0	1	0	0	71.4	0	1	4	3	1	7	1
1	0	1	1	0	0	68.3	1	1	4	4	1	7	1
1	0	0	1	0	0	58.0	1	1	4	4	1	7	1

Figure4.2 : Data set Sample in value label view

V. SYSTEM DESCRIPTIONS

What we want to do is to model a function – predict a loyal or not loyal student as a function of age, ch, sx, Rel, pob, etc. We are trying to build an adaptive system that will model an input-output relationship from our data file as Illustrated in figure 5.1

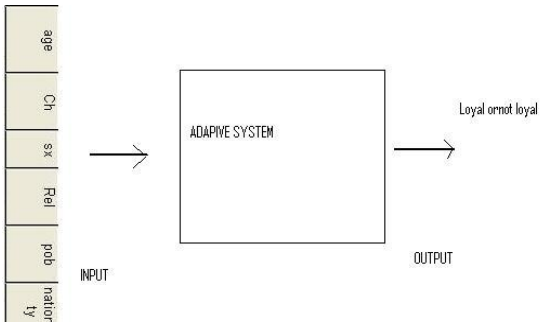


Figure5. 1: Adaptive System

We assigned 20% of data for validation; we did that because the Adaptive systems learn from data we supply it. But how can we be sure that they will work with other data – data that they haven't seen? The answer to that is validation. Instead of using all of the data available for training the system, we leave some aside with which we later test the system. This makes sure that we know how well the system is capable of generalization i.e. how well it works on data it hasn't been trained on. The 20% here refers to how much data should be put aside for validation figure 5.2.

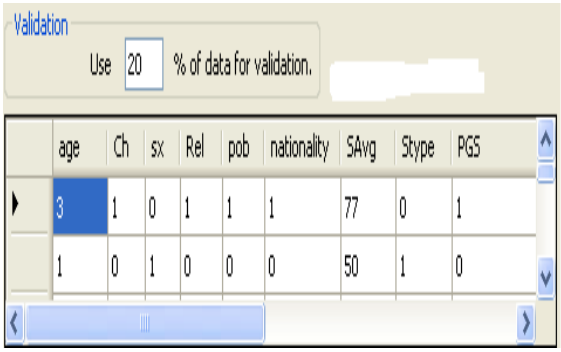


Figure5.2: assignment validation data

1) *Attribute Selection*

We needed to find a minimal set of attributes that preserve the class distribution. Attribute relevance is with respect to the class, i.e. relevant attribute and not relevant attribute. This will help us to select the best attributes from 14 attributes that have been collected. This will rank the attributes, according to the effectiveness of the features starting with the most significant feature, as follows in table 5.1.Before selecting the best attributes The figure 5.1.1 graphically depicts the instances in the dataset by using various combination of attributes as x/y axis values .

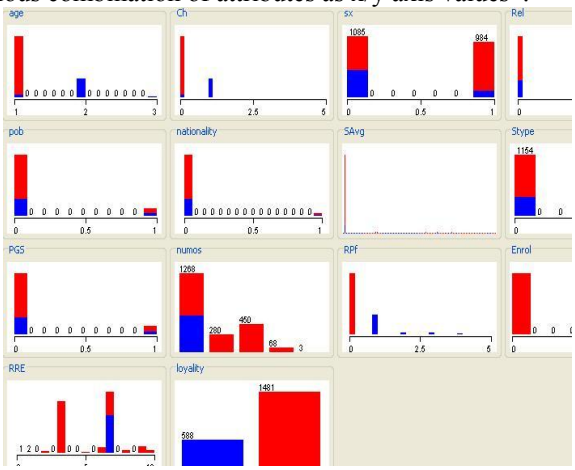


Figure 5.1.1 various combinations of attributes as x\y axis values.

Table 5.1.1: Ranked attributes, Attribute Selection on all input data

Instances: 2069	Evaluation mode: evaluate on all training data	Search Method: Best first.	Search direction: forward	Attribute Subset Evaluator (supervised, Class (nominal): 14 loyality)	Selected attributes: 1,4,11,12,13 age Rel RPf Enrol RRE

To select the best attributes we evaluated the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter correlation are preferred. For more information about calculation the worth of a subset of attributes in [17].For Search method we apply Best First, its Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility.

Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point).

2) Attribute Selection (ranking)

Our objectives are to build the classification model using a decision tree method. The decision tree is a very good method since it is relatively fast, it can be converted to simple rule, and accuracy level is high. The decision tree function in this project is focusing on classify the data in suites, to reach our data mining goal, which will be used to find the relationship between a specific features. By using the ID3 algorithm .In this project ID3 is applied on Vb.net programming language. The basic idea of ID3 algorithm is to construct the decision tree, by using the metric information gain, where attribute that is most useful, and by measuring which questions provide, the most balanced splitting the depth of the tree. The following rules will be applied with Vb.net language. Before discussing rules we used the metric information gain to reduce the entropy related to specified attribute when splitting decision tree node. The gain ratio evaluates the worth of an attribute by measuring the gain ratio with respect to the class.GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute).The Ranker ,Ranks attributes by their individual evaluations. Use in conjunction with attribute evaluators (ReliefF, GainRatio, Entropy etc). .We used the gain to rank attributes and to build our decision tree where each node is located the attribute with greatest gain, among the other attribute.

Table 5.2.1 , Attribute Selection (ranking)

Instances: 2069	Attributes: 14 Age, Ch, sx, Rel, pob, nationality, SAvg, Stype, PGS, numos, RPf, Enrol, RRE, loyality	Evaluation mode: evaluate on all training data	Search Method: Attribute ranking.	Attribute Evaluator (supervised, Class (nominal): 14 loyality): Gain Ratio feature evaluator	Ranked attributes: 1 11 RPf 1 12 Enrol 0.800821 age 0.800822 Ch 0.313071 3 RRE 0.249244 Rel 0.169571 0 numos 0.103213 sx 0.075517 SAvg 0.005465 pob 0 9 PGS 0 6 nationality 0 8 Stype	Selected attributes: 11,12,1,2,13,4,10,3,7,6,8:13

See figure below after filtering the selected attributes

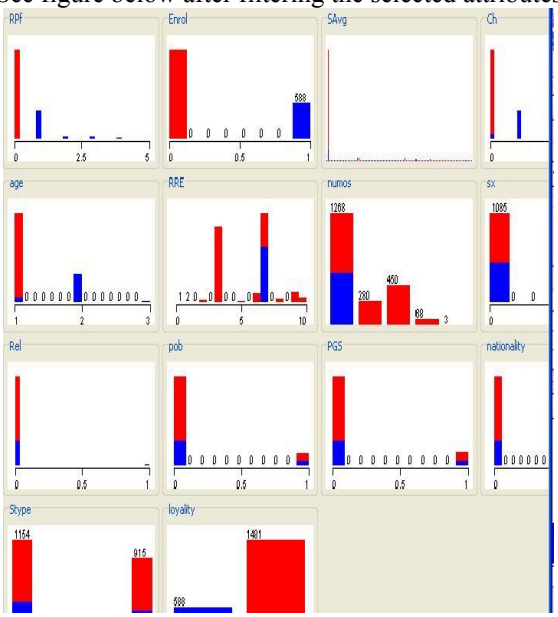


Figure 5.2.1 various combination of attributes as x\y axis values after filtering.

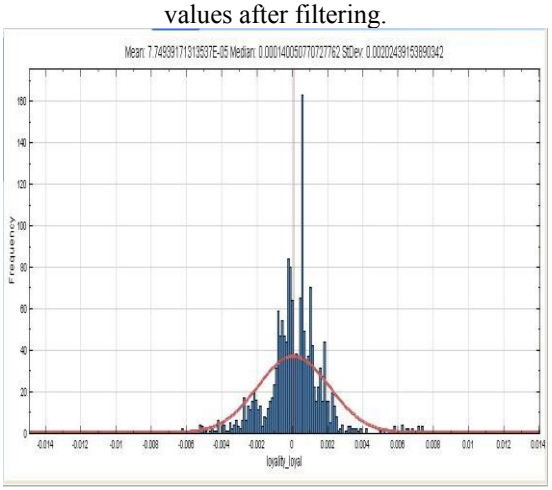


Figure 5.2.2 : Linear error analysis

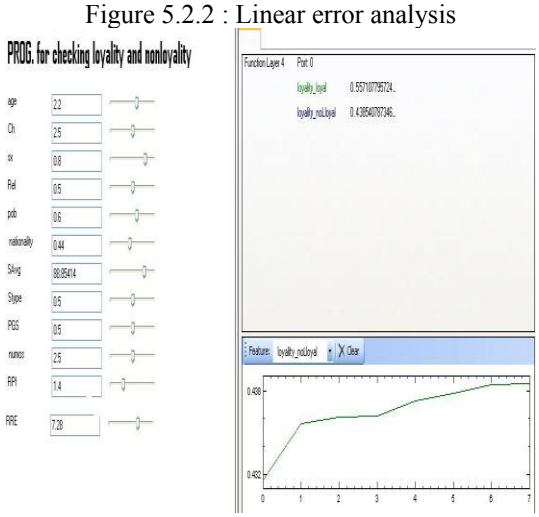


Figure 5.5.3: Sample of Program Interface

After providing the loyal, non loyal program with previous attributes of this program will Generate 551 rules, these rules will be used for predicting loyalty of enrolled students in a particular institute.Sample of generated rules:

IF (AGE =3 AND CH=1 AND SX = 0 AND REL= 1 AND POB = 1 AND NATIONALITY = 1 AND SAVG = 77 AND STYPE =0 AND PGS = 1 AND NUMOS = 4 AND RPF= 1 AND RPE = 7)THEN

Loyal studentElseNon loyal studentEnd if

age	Ch	sx	Rel	pob	nationality	SAvg	Stype	PGS	numos	RPf	Enrol	RRE	loyality
older	schild	Fem	other	forign	forign	77.0	litral	forign	4	1	Enr	much employe	Loy
yong	nochild	Male	Muslim	original	original	50.0	scintific	origina	4	0	not	near place	not
middle	schild	Fem	Muslim	original	original	59.8	scintific	origina	4	4	Enr	much employe	Loy
older	muchchi	Male	other	forign	forign	65.9	scintific	forign	4	4	Enr	much employe	Loy
yong	nochild	Fem	other	original	original	71.4	litral	forign	4	3	Enr	much employe	Loy
yong	nochild	Male	other	original	original	68.3	scintific	forign	4	4	Enr	much employe	Loy
yong	nochild	Fem	other	original	original	58.0	scintific	forign	4	4	Enr	much employe	Loy
middle	schild	Fem	other	original	original	89.0	litral	origina	4	2	Enr	much employe	Loy

VI. Conclusion

Data mining is a powerful analytical tool that enables educational institutions to better allocate resources and staff, and proactively manage student outcomes. With the ability to uncover hidden patterns in large databases, community colleges and universities can build models that predict with a high degree of accuracy the behavior of population clusters. By acting on these predictive models, educational institutions can effectively address issues ranging from transfers and retention, to marketing and alumni relations. The goal of education is to help people, especially young people, to participate in the functions of society, to acquire knowledge and to develop skills that will help them to confront the needs of the future and to be productive and competitive in tomorrow's world. This research paper is intended to enhance the quality of the higher educational system by focusing on using the data mining techniques. Using this research, the University will have the ability to predict the students loyalty (numbers of enrolled students) so they can manage and prepare necessary resources for the new enrolled students. Moreover, the higher managements can use the classification model to enhance the University resources according to the extracted knowledge. It is certain that the future holds a lot of surprises. It is another major task for education to prepare all resources that give young people the qualities and the skills for the jobs that do not exist yet and we believe that this research paper can help considerably towards that. It should also be a major task of the educational system to provide these qualities and skills in an enjoyable and modern way. The result of our experiment can be used to obtain a deep understanding of student's enrollment pattern in a University where the

faculty and managerial decision makers can utilize any action to provide extra basic course skill classes and academic counseling. In addition the management system can improve their policy, enhance their strategies and thereby improve the quality of that management system.

VII. References

1) Higher Education Enhancement Project (HEEP), 2007, http://www.heep.edu.eg.
2) Han J, Kamber M, Data Mining- Concepts and Techniques. Morgan KaufmannPublishers ,2001.
3) Luan J, "Data Mining and Its Applications in Higher Education" in A. Serban and J.
4) Luan (eds.) Knowledge Management: Building a Competitive Advantage for Higher Education. New Directions for Institutional Research, No. 113. San Francisco, CA: Jossey Bass (2002).
5) Nigel culkin, norbert morawetz, university of hertfordshire, centre for innovation and enterprise , www.hampp-verlag.de
   Hilbert, Andreas, Schnbrunn, Karoline, Schmode, Sophie."Student Relationship Management In Germany - Foundations And Opportunities", Management Revue, 2007.
6) http://www.britannica.com/bps/additionalcontent/18/25385295/Student-Relationship-Management-in-Germany--Foundations-and-Opportunities.
7) Delavari N, Beikzadeh M. R. A New Modfor Using Data Mining in Higher Educational System, 5th International Conference on Information Technology based Higher Education and Training: Istanbul, Turkey, May-2nd Jun 2004.
8) Mierle K, Laven K, Roweis S, Wilson G, Mining Student CVS Repositories for Performance Indicators, 2004.
9) Varapron P. et al. Using Rough Set theory for Automatic Data Analysis. 29th Congress on Science and Technology of Thailand. 2003.
10) Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery, Two Crows Corporation. Third Edition, U.S.A, 1999.
11) Han J, Kamber M. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, 2001.
12) Mehta M, Agrawal R, Rissanen J. SLIQ: A Fast Scalable Classifier for Data Mining in proc. 1996 int. conf. extending database technology (EDBT'96),Avignon, France, Mar 1996.
13) Murthy K. Automatic Construction Of Decision trees from Data : Multi-Disciplinary Survey. Siemens Corporate Research, Princeton, NJ 08540 USA.
14) Peng W,Chen J, Zhou H, An Implementation of ID-Decision Tree Learning Algorithem, Universty of New South Wales, Australia.
15) Esposito F, Malerba, D & Semeraro G, A Comparative Analysis of Methods for Pruning Decision Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 5, pp. 476-491, 1997.
16) Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R.
17) M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

Performance Evaluation of an Efficient Frequent Item sets-Based Text Clustering Approach

S.Murali Krishna ,S.Durga Bhavani

GJCST Classification (FOR)
H.2.8, I.7.m

***Abstract*-The vast amount of textual information available in electronic form is growing at a staggering rate in recent times. The task of mining useful or interesting frequent itemsets (words/terms) from very large text databases that are formed as a result of the increasing number of textual data still seems to be a quite challenging task. A great deal of attention in research community has been received by the use of such frequent itemsets for text clustering, because the dimensionality of the documents is drastically reduced by the mined frequent itemsets. Based on frequent itemsets, an efficient approach for text clustering has been devised. For mining the frequent itemsets, a renowned method, called Apriori algorithm has been used. Then, the documents are initially partitioned without overlapping by making use of mined frequent itemsets. Furthermore, by grouping the documents within the partition using derived keywords, the resultant clusters are obtained effectively. In this paper, we have presented an extensive analysis of frequent itemset-based text clustering approach for different real life datasets and the performance of the frequent itemset-based text clustering approach is evaluated with the help of evaluation measures such as, precision, recall and F-measure. The experimental results shows that the efficiency of the frequent itemset-based text clustering approach has been improved significantly for different real life datasets.**

Keywords-Text mining, Text clustering, Text documents, Frequent itemsets, Apriori, Reuter-21578, Webkb dataset, 20-newsgroups.

I. Introduction

Databases in every aspect of human actions progresses rapidly. This has led to an enormous demand for efficient tools that turn data into valuable knowledge. Researchers from numerous technological areas, namely, pattern recognition, machine learning, data visualization, statistical data analysis, neural networks, information retrieval, econometrics, information extraction etc have been searching for eminent approaches to fulfill this requirement. An effective research area known as data mining (DM) or Knowledge Discovery in Databases (KDD) has resulted as a result of their entire efforts [6]. Data represented in quantitative, multimedia or textual forms are commonly utilized for data mining [21]. Presently, in documents, news, email and manuals, large amount of information exists in the form of text. Because of the growth of digital libraries, web, technical documentation and medical data, access to a large quantity of textual documents turns out to be effectual. These textual data consists of resources which can be used in a better way. Text mining (TM) or in other words knowledge discovery from textual databases is a prominent and tough challenge in majority of the existing documents which employs natural language due to its value and ambiguity [1]. The need of acquiring knowledge fromlarge number of available text documents, particularly on the Web has made text mining as one of the major research fields [8].In some perspective, the information mining tasks such as, text mining and data mining are identical. Text can be mined by adapting the data mining techniques in a better way [8]. In both the knowledge-discovery variants, data is organized by tagging the document elements or by representing the numerical data in organized data structures. By identical application of statistical techniques, the complexity of the problem is lessened, correlations, linkages, clusters, and relationships are recognized, and predictive rules are made, which are also identified in few circles as knowledge [11, 12, and 13]. Text mining can be described as a knowledge-intensive process in which a user corresponds using a set of analysis tools in due course with a collection of documents. Similar to data mining by recognition and searching of interesting patterns, text mining anticipates valuable information from data sources. “Text mining” refers to the automated discovery of valuable or interesting information from unstructured text by the application of data mining techniques [4], [3], [2] and [10]. The requirement to attain knowledge from massive amount of text documents has led to a progressively more significant research field known as text mining [34]. Pre-processing of text documents and subsequent saving of data in a data structure that is more suitable for further processing than a plain text file are essential to mine large document collections [18]. Typically, tokenization, Part of Speech (PoS) Tagging [19], word stemming and the application of a stop words removal technique are included in text preprocessing. The process of splitting the text into words or terms is known as tokenization. According to the grammatical context of a word in a sentence, it can be tagged using Part of Speech Tagging and thereby words can be divided into nouns, verbs and more [20]. Extract of key information from the original text represented by filling predefined structured representation, or templates by mapping natural language texts (namely newswire reports, journal and newspaper articles, World Wide Web pages, electronic mail, any textual database and more.) is defined as Information Extraction [7]. Lately, significant interest has been achieved by extracting relationships from entities in text documents. The interesting associations and/or correlation relationships among large set of data items are discovered by association rule mining.In a wide range of application areas, mining association rules in transaction databases has been demonstrated to be valuable [15, 16]. Due to the difference in characteristics between transaction databases and text databases, application of association rules mining seems to be

About-Associate Professor Department of Computer Science and Engineering
smuralikrishnaphd@gmail.com

more promising in text databases [14]. In the case of text mining, extracted rules are able to return semantic relations among the terms because they are deduced as co-occurrences of terms in texts [17]. Text analysis, information retrieval, clustering, information extraction, categorization, visualization, machine learning, data mining and database technology are functions included in the multidisciplinary field of text mining [22]. Cluster analysis is the method of dividing data objects (e.g.: document and records) into significant clusters or groups such that contradictory to objects in other clusters, analogous characteristics are possessed by objects within a cluster [4], [5]. Navigation, summarization, and arrangement of text documents by users are assisted by text clustering which sorts several documents in an efficient way [31, 32, 9]. Document clustering, which arranges a huge number of documents into meaningful clusters, can be employed to browse a set of documents or to arrange the results given by a search engine in answer to a user's query. The accuracy and recall in information retrieval systems can be considerably enhanced and the nearest neighbors of a document can be determined in a proficient way by document clustering [33]. In our prior research [45], we have presented an effective frequent itemset-based document clustering approach. First, with the aid of stop words removal technique and stemming algorithm, the text documents in the text data are preprocessed. Then, from each document, the top-p frequent words are extracted and the binary mapped database is formed using the extracted words. The different length frequent itemsets are discovered by applying the Apriori algorithm. Based on the support level of the mined frequent itemsets, the itemsets are sorted in descending order for each length. Subsequently, using the sorted frequent itemsets, we split the documents into partition. Understandable description of the partitions can be obtained from these frequent itemsets. Furthermore, the derived keywords (words obtained by taking the absolute complement of familiar keywords with respect to the top-p frequent words) are used to form the resultant cluster within the partition. In this paper, we have used the different real life datasets such as, Reuter-21578, 20-newsgroups and Webkb datasets for analysing the frequent itemset-based document clustering approach. In addition with, for evaluating the performance, we make use of evaluation metrics namely, precision, recall and F-measure.The paper is organized as follows. The concise review of related researches of text clustering is presented in Section 2. The text clustering approach based on frequent itemset is described in section 3. The extensive analysis of the text clustering approach using different datasets is given in section 4. The conclusion is summed up in section 5.

II. Literature Review of Recent Related Researches

The literature presents a lot of text clustering approach, wherein the frequent itemset based text clustering has received considerable attention among the research community. Some recent researches related to frequent itemset-based text clustering is briefly reviewed in this section.Zhou Chong *et al.* [23] have presented a method for text clustering known as Frequent Itemset-based Clustering with Window (FICW), in which the semantic information has been employed with a window constraint. FICW has revealed better performance in terms of clustering accuracy and efficiency in the experimental results obtained from three (hypertext) text sets. Xiangwei Liu and Pilian [24] have introduced a text-clustering algorithm known as Frequent Term Set-based Clustering (FTSC) which clusters texts by employing frequent term sets. Initially, significant information from the documents are extracted by it and kept in databases. Later, the frequent item sets are mined by it employing the Apriori algorithm. Finally, the documents are clustered as per the frequent words in subsets of the frequent term sets. The dimension of the text data can be lessened by the algorithm for extremely large databases, so the accuracy and speed of the clustering algorithm can be enhanced. Experimental results have showed that the clustering performance efficiency of FTSC and FTSHC algorithms are superior to that of the K-Means algorithmLe Wang *et al.* [25] have presented a top-k frequent term sets and k-means based simple hybrid algorithm (SHDC) to overcome the main challenges of current web document clustering. K initial means regarded as initial clusters were provided by employing top-k frequent term sets, which were later refined by k-means. K-means returns the final optimal clustering whereas k frequent term sets provides the clear description of clustering. Efficiency and effectiveness of SHDC was proved to be superior to that of the other two representative clustering algorithms (the farthest first k-means and random initial k-means) by the experimental results conducted on two public datasets. Zhitong Su *et al.* [26] have introduced a maximal frequent itemsets based web-text clustering method for personalized e-learning. The Web documents were initially represented by vector space model. Maximal frequent word sets were determined subsequently. In the end, documents were clustered by employing maximal itemsets on the basis of a new similarity measure of itemsets. The method was proved to be efficient by the obtained Experimental results.Yongheng Wang *et al.* [27] have introduced a parallel clustering algorithm based on frequent term for clustering short documents in very large text database. To enhance the clustering accuracy, a semantic classification method has also been employed. The algorithm has been proved to be more precise and efficient than other clustering algorithms when clustering large scale short documents based on experimental analysis. In addition, the scalability of the algorithm is good and also huge data could be processed by employing it. W.L. Liu and X. S. Zheng have proposed the frequent term sets based documents clustering algorithm [29]. Initially, by means of the Vector Space Model (VSM), the documents were denoted and all the terms were sorted in accordance with their relative frequency. Then, frequent-pattern growth (FP growth) has been used to mine the frequent term sets. Lastly, on the basis frequent term sets the documents were clustered. The approach has been efficient in very large databases and also a clear explanation in terms of frequent terms about the determined clusters has been provided by the algorithm. With the aid of experimental results, the efficiency and suitability of the proposed algorithm has been demonstrated.Henry Anaya-Sanchez *et al.* [30] have

proposed a text clustering algorithm which depending on the most probable term pairs of the collection and its estimate of related topic homogeneity, discovers and unfolds the topics, included in the text collection. From term pairs that have support sets with sufficient homogeneity for denoting collection topics, topics and their descriptions were produced. The efficacy and usefulness of the approach has been verified by the experimental results obtained over three benchmark text collections. Florian Beil *et al.* [28] have proposed an approach for text clustering which employs frequent item (term) sets. Utilizing algorithms for association rule mining such frequent sets were determined. They have gauged the mutual overlap of frequent sets with regard to the sets of supporting documents to cluster on the basis of frequent term sets. FTC and HFTC are the two algorithms, which they have provided for frequent term-based text clustering. Flat clustering is produced by FTC whereas HFTC is for obtaining hierarchical clustering. Clustering of the presented algorithm has been proved to be of comparable quality and appreciably more efficiency than modern text clustering algorithms based on an experimental assessment on classical text as well as web documents.

III. Proficient Approach For Text Clustering Based On Frequent Itemsets

The exploration for hidden knowledge in text collections has been provoked by the reputation of the Web and the huge quantity of documents existing in electronic form. Therefore, research concentration is increasing in the general topic of text mining. One of the popular techniques among the various data mining techniques that have been used by researchers for finding the meaningful information from the text documents is clustering. Grouping a collection
of documents (unstructured texts) into different category groups so that the same subject is described by documents in the same category group is known as text clustering. Possible ways to improve the performance of text document clustering based on the popular clustering algorithms (partitional and hierarchical clustering) and frequent term based clustering has been investigated by many researches [23-26, 28]. An effectual approach for clustering a text corpus with the aid of frequent itemsets [45] is discussed in this section. The text clustering approach consists of the following major steps:

1) Text preprocessing
2) Mining of frequent itemsets
3) Partitioning the text documents based on frequent Itemsets
4) Clustering of text documents within the partition

1) *Text Pre-processing*

Let D be a set of text documents represented as $D = \{d_1\ d_2\ d_3 \ldots . d_n\}; 1 \le i \le n$, where, n is the number documents in the text dataset D. The words or terms are extracted (tokenization) from the text document set D using the text preprocessing techniques and it is converted from unstructured format into some common representation. For preprocessing the input data set D (text documents), two techniques namely, removing stop words and stemming algorithm are used.

(a) Stop word Removal: Stop (linking) words like “have”, “then”, "it", "can”, “need", "but", "they", "from", "was", "the", "to", "also" are removed from the document [36].

(b) Stemming algorithm: Prefixes and suffixes of each word [35] are removed.

2) *Mining of Frequent Itemsets*

Mining of frequent itemsets from the preprocessed text documents D is described in this sub-section. After the preprocessing step, the frequency of the extracted words or terms is computed for every document d_i, and the top- p frequent words from each document d_i are taken out.

$$K_w = \{d_i \mid \mathrm{p}(\mathrm{d_i}) \quad ; \quad \forall d_i \subseteq D\}$$

where, $p(d_i) = T_{w_j} \quad ; \quad 1 \le j \le p$

By using the unique words of the set of top- p frequent words, the binary database B_T is formed. Let B_T be a binary database consisting of n number of transactions (documents) T and q number of attributes (unique words) $U = [u_1, u_2, \ldots, u_q]$. Whether the unique words are present or not in the documents, is represented in the binary database B_T, which consists of binary data.

$$B_T = \begin{cases} 0 & if \quad u_j \notin d_i \\ 1 & if \quad u_j \in d_i \end{cases}$$

; $1 \le j \le q$, $1 \le i \le n$

Then, the frequent itemsets (words/terms) F_s is mined by inputting the binary database B_T to the Apriori algorithm.

a) *Apriori algorithm*

Apriori algorithm, first introduced in [37] is a conventional algorithm for mining association rules. Association rules mining involves two steps such as, (1) Identifying frequent itemsets (2) Generating association rules from the frequent itemsets. Two steps are involved in frequent itemsets mining. Generating the candidate itemsets is the first step. In the second step, assisted by these candidate itemsets, the frequent itemsets are mined. Frequent itemsets consists of itemsets whose support is greater than the user specified minimum support.In our proposed approach, we use only the frequent itemsets for further processing. The pseudo code corresponding to first step (generation of frequent itemsets) of the Apriori algorithm [38] is given below.

Pseudo code:

C_k : Candidate itemset of size k

I_k : Frequent itemset of size k.

$I_1 = \{large\ 1-itemsets\};$
for $(k = 2;\ I_{k-1} \neq 0;\ k++)$ *do begin*
$C_k = apriori-gen(I_{k-1});$ *// New candidates*
for all transactions $T \in D$ *do begin*
$C_T = subset(C_k, T);$
// Candidates contained in T
for all candidates $c \in C_T$ *do*
$c.count++;$
end
end
$I_k = \{c \in C_k \mid c.count \geq minsup\}$
end
$Answer = \bigcup_k I_k;$

C. Partitioning the Text Documents Based on Frequent Itemsets

Partitioning of text documents (D) based on mined frequent itemsets (F) is described in this section.

Definition1:*Frequent itemset* is a set of words that occur together in some minimum fraction of documents in a cluster. A set of frequent itemsets of varying length (l), from 1 to k, are generated by the Apriori algorithm. First, the set of frequent itemsets of each length (l) are sorted in descending order of their support level.

$$F_s = \{ f_1\ f_2\ f_3 \ldots . f_k \}\ ;\ 1 \leq l \leq k$$
$$f_l = \{ f_{l(i)}\ ;\ 1 \leq i \leq t \}$$

where, $\sup(f_{l(1)}) \geq \sup(f_{l(2)}) \geq \ldots \geq \sup(f_{l(t)})$ and t denotes the number of frequent itemsets in the set f_l.At first, the first element ($f_{(k/2)}(1)$) is selected from the sorted list $f_{(k/2)}$, that contains a set of frequent itemsets. Then, an initial partition c_1 is constructed by grouping all the documents that contains the itemset $f_{(k/2)}(1)$. After that, a new partition c_2 is formed for the second element $f_{(k/2)}(2)$, the support of which is less than $f_{(k/2)}(1)$. This new partition c_2 is formed by grouping all the documents that have the frequent itemset $f_{(k/2)}(2)$ and also, the documents in the initial partition c_1 are taken away from this new partition. This process is done repeatedly until every text documents in the input dataset D are moved into partition $C_{(i)}$. In addition, if the above procedure is not terminated with the sorted list $f_{(k/2)}$, then the above discussed steps (inserting the documents into the partition) are performed by taking the subsequent sorted lists ($f_{((k/2)-1)}$, $f_{((k/2)-2)}$ etc..). This provides a set of partition c and each partition $C_{(i)}$ contains a collection documents $D^{(x)}_{c(i)}$.

$$c = \{ c_{(i)} \mid c_{(i)} \in f_{l(i)} \}\ \ ;\ \ 1 \leq i \leq m\ ,\ \ 1 \leq l \leq k$$
$$C_{(i)} = Doc[f_{l(i)}]$$
$$C_{(i)} = \{ D^{(x)}_{c(i)}\ ;\ D^{(x)}_{c(i)} \in D\ ,\ 1 \leq x \leq r \}$$

where, m denotes the number of partitions and r indicates the number of documents in each partition.

Mined frequent itemset used for constructing initial partition (or cluster), significantly reduces the dimensionality of the text document set, and the reduced dimensionality considerably enhances the efficiency and scalability of clustering. Due to the use of frequent itemsets, overlapping of documents exists in the clustering results produced by the approaches presented in [41, 28] and the final results are obtained by removing these overlapping documents. In the proposed research, non-overlapping partitions are directly generated from the frequent itemsets. This makes the initial partitions disjoint, because the document is kept only within the best initial partition by this text clustering approach.

3) Clustering of Text Documents within the Partition

This sub-section, describes how clustering is done on the set of partitions obtained from the previous step. This step is necessary to form a sub cluster (describing sub-topic) of the partition (describing same topic) and the outlier documents can be significantly detected by the resulting cluster.Furthermore, a pre-specified number of clusters are not required by this text clustering approach. The devised procedure for clustering the text documents available in the set of partition c is discussed below.

In this phase, for each document $D^{(x)}_{c(i)}$, the familiar words $f_{c(i)}$ (frequent itemset used for constructing the partition) of each partition $C_{(i)}$ are first identified. Then, by taking the absolute complement of familiar words $f_{c(i)}$ with respect to the top- p frequent words of the document, the derived keywords $K_d[D^{(x)}_{c(i)}]$ of document $D^{(x)}_{c(i)}$ are obtained.

$$K_d[D^{(x)}_{c(i)}] = \{ T_{w_j} \setminus f_{c(i)} \}\ ;\ T_{w_j} \in D^{(x)}_{c(i)}\ \ ,$$
$$1 \leq i \leq m\ \ ,\ 1 \leq j \leq p\ \ ,1 \leq x \leq r$$
$$T_{w_j} \setminus f_{c(i)} = \{ x \in T_{w_j} \mid x \notin f_{c(i)} \}$$

For each partition $C_{(i)}$, the set of unique derived keywords and their support are computed within the partition. The representative words of the partition $C_{(i)}$ are formed by the set of keywords which satisfy the cluster support ($cl\_\sup$). Definition2: The percentage of the documents in $C_{(i)}$ that contains a keyword is the *cluster support* of that keyword in $C_{(i)}$.

$$R_w[\, c(i)\,] = \{ x : p(x) \}$$

where, $p(x) = [K_d[D^{(x)}_{c(i)}]] \geq cl\_\sup$

Subsequently, with respect to the representative words $R_w[c(i)]$, the similarity of the documents $D^{(x)}_{c(i)}$ is computed. An important role is played by the definition of similarity measure in obtaining effective and meaningful clusters. The similarity between two text documents S_m is calculated as follows,

$$S\left(K_d[D^{(x)}_{c(i)}], R_w[c(i)]\right) = \left|K_d[D^{(x)}_{c(i)}] \cap R_w[c(i)]\right|$$

$$S_m\left(K_d[D^{(x)}_{c(i)}], R_w[c(i)]\right) = \frac{S\left(K_d[D^{(x)}_{c(i)}], R_w[c(i)]\right)}{\left|R_w[c(i)]\right|}$$

2. Experimentation And Performance Evaluation

This section details the experimentation and performance evaluation of the frequent itemset-based text clustering approach. We have implemented the frequent itemset-based text clustering approach using Java (JDK 1.6). The text clustering approach is evaluated based on the evaluation metrics given in sub-section 4.1 and in sub-section 4.2, the performance of the text clustering approach is analyzed with the different real life datasets.

1) Evaluation measures

Precision, Recall and F-measure described in [39, 40] are used for evaluating the performance of the frequent itemset-based text clustering approach. The definition of the evaluation metrics is given below,

$$\text{Precision}(i,j) = M_{ij} / M_j$$

$$\text{Recall}(i,j) = M_{ij} / M_i$$

$$\textbf{F-measure}\ (i,j) = \frac{2 * \text{Recall}\ (i,j) * \text{Precision}\ (i,j)}{\text{Precision}\ (i,j) + \text{Recall}\ (i,j)}$$

where M_{ij} is the number of members of topic i in cluster j, M_j is the number of members of cluster j and M_i is the number of members of topic i.

2) Performance Evaluation

The experimentation is carried out on different datasets based on the steps described in the section 3. At first, the top- p frequent words are extracted from each document and binary database is constructed using these frequent words. Using Apriori algorithm, frequent itemsets are mined from the binary database and sorted it based on their support level. Then, we construct initial partition using these frequent itemsets. Subsequently, we compute representative words of each partition with the help of top- p frequent words and familiar words. For each document, we calculate the similarity measure with respect to the representative words. Finally, if the similarity value of the document within the partition is below 0.4, it forms as a separate cluster. The detailed evaluation of the frequent itemset-based text clustering approach for different real life datasets is described below.

Dataset 1: We have taken 100 documents manually from various topics namely, Content based video retrieval, Semantic web, Incremental clustering, Gene prediction, Human Resource, Sequential pattern mining, Adaptive e-learning, Multimodal Biometrics, Public key cryptography, Automatic text summarization and Grid Computing. These documents are fed as an input to the text clustering approach that provides 25 clusters. The performance of the resulted cluster is evaluated with three measures (Precision, Recall and F-measure) and the obtained result is given in table 1. The plotted graph for the dataset 1 is given in figure 1. By analyzing the graph shown in figure 1, some of the resulted cluster has achieved maximum precision.

Table 1. Precision, Recall and F-measure of dataset 1

Partition	Cluster	Precision	Recall	F-measure
P_1	C_1	1	0.666667	0.8
	C_2	0.454545	0.714286	0.555555
P_2	C_3	1	0.4	0.571429
	C_4	0.222222	0.2	0.210526
P_3	C_5	1	0.5	0.666667
	C_6	0.375	0.428571	0.4
P_4	C_7	0.666667	0.2	0.307692
	C_8	0.571429	0.4	0.470588
P_5	C_9	1	0.2	0.333333
	C_{10}	0.25	0.142857	0.181818
P_6	C_{11}	1	0.2	0.333333
	C_{12}	0.333333	0.125	0.181818
P_7	C_{13}	1	0.555556	0.714286
	C_{14}	0.4	0.2	0.266667
P_8	C_{15}	0.75	0.333333	0.461538
	C_{16}	0.333333	0.142857	0.2
P_9	C_{17}	1	0.25	0.4
	C_{18}	0.666667	0.2	0.307692
P_{10}	C_{19}	1	0.222222	0.363636
	C_{20}	0.5	0.125	0.2
P_{11}	C_{21}	1	0.2	0.333333
P_{12}	C_{22}	1	0.222222	0.363636
	C_{23}	1	0.111111	0.2
P_{13}	C_{24}	1	0.2	0.333333
P_{14}	C_{25}	1	0.25	0.4

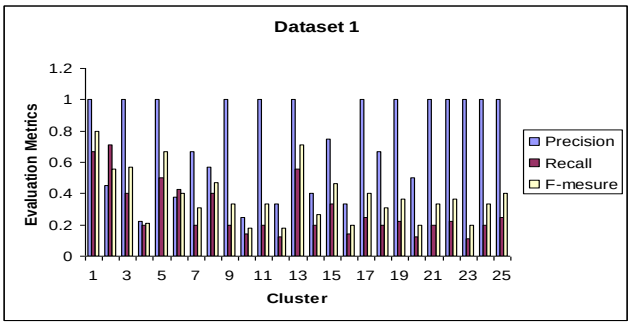


Fig.1. Performance of the frequent itemset-based text clustering approach on dataset 1

(2) Reuter 21578 dataset: The documents in the Reuters-21578 set [42] resembled on the Reuters newswire in 1987. The documents were accumulated and indexed with grouping, by personnel from Reuters Ltd. Additionally, formatting and data file production was achieved in 1991 and 1992 by David D. Lewis and Peter Shoemaker at the Center for Information and Language Studies, University of Chicago. For experimentation, we have taken 125 documents from 10 different topics (cpi, bop, cocoa, coffee, crude, earn, trade, acq, money-fx, oilseed) and these documents is given as input documents to the text clustering approach. It provides 24 clusters and for each cluster, the precision, Recall and F-measure is computed. The results obtained are given in table 2 and their corresponding graph is shown in figure 2. It ensures that some of the clusters obtained its maximum precision and recall measures.

Table 2. Precision, Recall and F-measure of Reuter 21578

Partition	Cluster	Precision	Recall	F-measure
P_1	C_1	0.8	0.5333	0.639976
	C_2	0.2857	0.5454	0.374975
P_2	C_3	1	0.5	0.666667
	C_4	0.9285	1	0.962925
P_3	C_5	1	0.4166	0.588169
	C_6	0.4444	0.2666	0.333269
P_4	C_7	1	0.2666	0.42097
	C_8	1	0.3333	0.499962
P_5	C_9	0.6666	0.1333	0.222172
P_6	C_{10}	1	0.2857	0.444427
	C_{11}	0.75	0.25	0.375
P_7	C_{12}	1	0.2857	0.444427
	C_{13}	0.5	0.0909	0.153833
P_8	C_{14}	1	0.2727	0.428538
	C_{15}	0.3333	0.0909	0.142843
P_9	C_{16}	1	0.0666	0.124883
P_{10}	C_{17}	0.75	0.2	0.315789
P_{11}	C_{18}	0.5	0.0833	0.142808
P_{12}	C_{19}	1	0.2727	0.428538
	C_{20}	0.6	0.2727	0.374974
P_{13}	C_{21}	1	0.5	0.666667
	C_{22}	0.5	0.25	0.333333
P_{14}	C_{23}	1	0.0714	0.133284
P_{15}	C_{24}	0.75	0.25	0.375

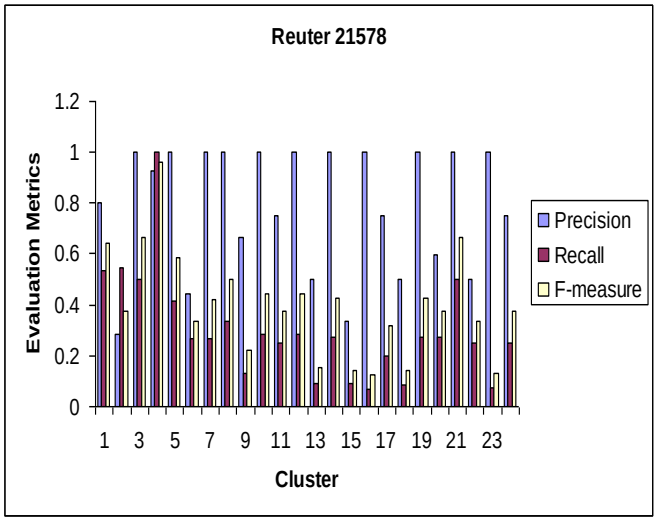


Fig.2. Performance of the frequent itemset-based text clustering approach on Reuter-21578

(3) 20 newsgroups dataset: This data set (20NG) [43] contains 20000 messages obtained from 20 newsgroups and 1000 messages are collected from each newsgroup. The various newsgroups prescribed in the dataset are alt.atheism, comp.graphics,comp.os.mswindows.misc,comp.sys.ibm.pc.hardware,comp.sys.mac.hardware,comp.windows.x, misc.forsale, rec.autos,rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt,sci.electronics,sci.med,sci.space,soc.religion.christian ,talk.politics.guns,talk.politics.mideast,talk.politics.misc,talk.religion.misc. For evaluation, the text clustering approach is applied on 201 documents taken from various newsgroups of 20NG dataset. Consequently, we have obtained 29 clusters and the resulted cluster is used to find the precision, recall and f-measure. The measured parameter is given in table 3 and the graph shown in figure 3 is plotted based on the measured parameters.

Table 3. Precision, Recall and F-measure of 20-newsgroups

Partition	Cluster	Precision	Recall	F-measure
P_1	C_1	1	0.565217	0.722222
P_2	C_2	1	0.578947	0.733333
P_3	C_3	0.875	0.233333	0.368421
P_4	C_4	0.5	0.333333	0.4
P_5	C_5	0.428571	0.130435	0.2
P_6	C_6	1	0.263158	0.416667
P_7	C_7	0.5	0.130435	0.206897
P_8	C_8	0.857143	0.24	0.375
P_9	C_9	1	0.466667	0.636364
P_{10}	C_{10}	0.857143	0.2	0.324324
P_{11}	C_{11}	0.666667	0.235294	0.347826
P_{12}	C_{12}	1	0.083333	0.153846
	C_{13}	1	0.166667	0.285714
P_{13}	C_{14}	0.5	0.181818	0.266667
P_{14}	C_{15}	1	0.352941	0.521739
P_{15}	C_{16}	1	0.24	0.387097
P_{16}	C_{17}	0.5	0.066667	0.117647
P_{17}	C_{18}	1	0.384615	0.555556
P_{18}	C_{19}	1	0.416667	0.588235
P_{19}	C_{20}	0.5	0.166667	0.25
P_{20}	C_{21}	1	0.166667	0.285714
	C_{22}	0.368421	0.233333	0.285714
P_{21}	C_{23}	0.333333	0.153846	0.210526
P_{22}	C_{24}	0.333333	0.307692	0.32
P_{23}	C_{25}	0.2	0.066667	0.1
P_{24}	C_{26}	0.25	0.25	0.25
P_{25}	C_{27}	0.4	0.5	0.44444
P_{26}	C_{28}	0.75	0.24	0.363636
P_{27}	C_{29}	0.25	0.1	0.142857

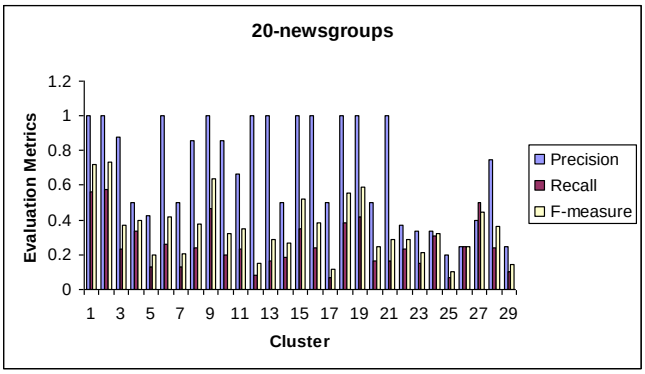


Fig.3. Performance of the frequent itemset-based text clustering approach for 20-newsgroups

(4) Webkb dataset: This data set [44] consists of WWW-pages collected from computer science departments of different universities namely, Cornell, Texas, Misc, Washington, Wisconsin in January 1997 by the World Wide Knowledge Base (Web->Kb) project of the CMU text learning group. The 8,282 pages were manually categorized into the following categories: student (1641), faculty (1124), staff (137), department (182), course (930), project (504) and other (3764). In order to evaluate the text clustering approach on Webkb dataset, we have taken 395 documents from various topics. These documents are used as input text documents and lastly, it results 27 clusters. We compute the precision, Recall and F-measure for each cluster and the attained parameters are shown in table 4. The graph for the results is given in figure 4. The obtained results show the efficiency of the text clustering approach.

Table 4. Precision, Recall and F-measure of WebKb dataset

Partition	Cluster	Precision	Recall	F-measure
P_1	C_1	0.84415584	0.77844311	0.80996885
P_2	C_2	0.67088608	0.37857143	0.48401826
P_3	C_3	0.60869565	0.1	0.17177914
	C_4	0.31818182	0.53846154	0.4
P_4	C_5	1	0.07857143	0.14569536
	C_6	0.6	0.04285714	0.08
P_5	C_7	0.82352941	0.1	0.17834395
P_6	C_8	1	0.18181818	0.30769231
	C_9	0.625	0.03571429	0.06756757
P_7	C_{10}	1	0.02142857	0.04195804
	C_{11}	0.5	0.06060606	0.10810811
P_8	C_{12}	0.75	0.01796407	0.03508772
	C_{13}	0.75	0.01796407	0.03508772
P_9	C_{14}	0.33333333	0.07692308	0.125
P_{10}	C_{15}	1	0.02142857	0.04195804
P_{11}	C_{16}	1	0.01197605	0.02366864
	C_{17}	0.5	0.15384615	0.23529412
P_{12}	C_{18}	1	0.01197605	0.02366864
	C_{19}	0.75	0.01796407	0.03508772
P_{13}	C_{20}	1	0.02142857	0.04195804
	C_{21}	1	0.07692308	0.14285714
P_{14}	C_{22}	0.5	0.00598802	0.01183432
P_{15}	C_{23}	0.5	0.00714286	0.01408451
P_{16}	C_{24}	0.33333333	0.03448276	0.0625
P_{17}	C_{25}	0.5	0.00598802	0.01183432
P_{18}	C_{26}	1	0.01197605	0.02366864
P_{19}	C_{27}	0.47058824	0.04790419	0.08695652

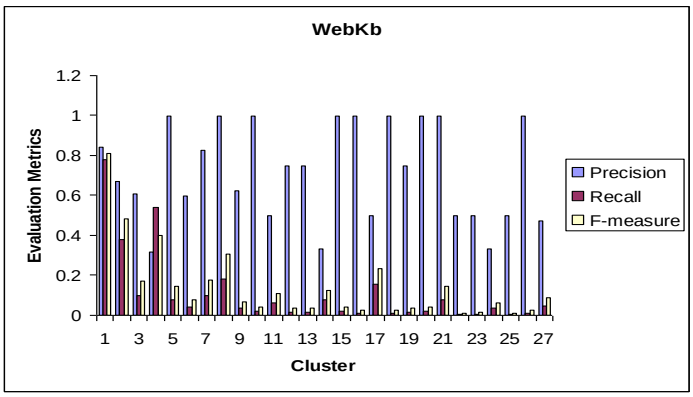


Fig.4. Performance of the frequent itemset-based text clustering approach on WebKb dataset

V. Conclusion

Text clustering is a more specific method for unsupervised text grouping, automatic topic extraction and fast information retrieval or filtering. There has been a plenty of approaches available in the literature for clustering the text documents. In this paper, we have conducted an extensive analysis of frequent itemset-based text clustering approach with different text datasets. For different text datasets, the performance of frequent itemset-based text clustering approach has been evaluated with precision, recall and F-measure. The experimental results of the frequent itemset-based text clustering approach are given for Reuter-21578, 20-newsgroups and Webkb datasets. The performance study of the text clustering approach showed that it effectively groups the documents into cluster and mostly, it provides better precision for all datasets taken for experimentation.

VI. References

1) Hany Mahgoub, Dietmar Rosner, Nabil Ismail and Fawzy Torkey, "A Text Mining Technique Using Association Rules Extraction", International Journal of Computational Intelligence, Vol. 4; No. 1, 2008.
2) Shenzhi Li, Tianhao Wu, William M. Pottenger, "Distributed Higher Order Association Rule Mining Using Information Extracted from Textual Data", ACM SIGKDD Explorations Newsletter, Natural language processing and text mining Vol. 7, No. 1 , pp. 26 - 35 , 2005.
3) R. Baeza-Yates, B. Ribeiro-Neto. "Modern Information Retrieval", ACM Press, New York, 1999.
4) J. Han, M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco, 2000.
5) Jochen Dijrre, Peter Gerstl, Roland Seiffert, "Text Mining: Finding Nuggets in Mountains of Textual Data", Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining , San Diego, California, United States , pp: 398 - 401, 1999.
6) Haralampos Karanikas, Christos Tjortjis and Babis Theodoulidis, "An Approach to Text Mining using Information Extraction", Proc. Knowledge Management Theory Applications Workshop, (KMTA 2000), Lyon, France, pp: 165-178, September 2000.
7) Wilks Yorick, "Information Extraction as a Core Language Technology", International Summer School, SCIE-97, 1997.
8) Ah-hwee Tan, "Text Mining: The state of the art and the challenges", In Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases, pp. 65-70,1999.
9) Jain, A.K., Murty, M.N., Flynn, P.J., "Data Clustering: A Review", ACM Computing Surveys, Vol: 31, No: 3, pp: 264-323. 1999.
10) Feldman, R., Sanger, J., "The Text Mining Handbook", Cambridge University Press, 2007.
11) Seth Grimes, "The Developing Text Mining Market", White paper from Alta Plana Corporation, Text Mining Summit, 2005.
12) M. Grobelnik, D. Mladenic, and N. Milic-Frayling, "Text Mining as Integration of Several Related Research Areas: Report on KDD'2000 Workshop on Text Mining," 2000.
13) M. Hearst, "Untangling Text Data Mining," in the Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999.
14) Alisa Kongthon, "A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management", Technical Report, Georgia Institute of Technology, April 2004.
15) Brijs, Tom, Swinnen, Gilbert, Vanhoof, Koen and Wets, "Using Association Rules for Product Assortment Decisions: A Case Study", In proceedings of Knowledge Discovery and Data Mining, pp: 254–260, 1999.
16) Dong, Jianning, Perrizo, William, Ding, Qin and Zhou, "The Application of Association Rule Mining to Remotely Sensed Data", In proceedings of the ACM symposium on Applied computing, Vol.1, pp: 340–345, 2000.
17) Valentina Ceausu and Sylvie Despres, "Text Mining Supported Terminology Construction", In proceedings of the 5th International Conference on Knowledge Management, Graz, Austria, 2005.
18) Hotho, Nurnberger and Paass, "A Brief Survey of Text Mining Export", LDV Forum, Vol.20, No.2, pp.19-62, 2005.
19) Manning and Schütze, "Foundations of statistical natural language processing", MIT Press, 1999.
20) Shatkay and Feldman, "Mining the Biomedical Literature in the Genomic Era: An Overview", Journal of Computational Biology, Vol.10, No.6, pp.821-855, 2003.
21) Pegah Falinouss, "Stock Trend Prediction using News Articles", Technical Report, Lulea. University of Technology, 2007.
22) Nasukawa and Nagano, "Text Analysis and Knowledge Mining System", IBM Systems Journal, Vol.40, No.4, pp.967-984, October 2001.
23) Zhou Chong, Lu Yansheng, Zou Lei and Hu Rong, "FICW: Frequent itemset based text clustering with window constraint", Wuhan University Journal of Natural Sciences, Vol: 11, No: 5, pp: 1345-1351, 2006.
24) Xiangwei Liu and Pilian He, "A Study on Text Clustering Algorithms Based on Frequent Term Sets", Lecture Notes in Computer Science, Vol:3584,pp:347-354, 2005.
25) Le Wang, Li Tian, Yan Jia and Weihong Han, "A Hybrid Algorithm for Web Document Clustering Based on Frequent Term Sets and k-Means", Lecture Notes in Computer Science, Springer Berlin ,Vol: 4537, pp: 198-203, 2010.
26) Zhitong Su ,Wei Song ,Manshan Lin ,Jinhong Li, "Web Text Clustering for Personalized E-learning

Based on Maximal Frequent Itemsets", Proceedings of the 2008 International Conference on Computer Science and Software Engineering , Vol: 06, Pages: 452-455 , 2008.

27) Yongheng Wang , Yan Jia and Shuqiang Yang, "Short Documents Clustering in Very Large Text Databases", Lecture Notes in Computer Science, Springer Berlin ,Vol:4256, pp: 83-93, 2006.
28) Florian Beil, Martin Ester and Xiaowei Xu, " Frequent term-based text clustering", in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada, pp. 436 - 442 , 2002.
29) W.-L. Liu and X.-S. Zheng, "Documents Clustering based on Frequent Term Sets", Intelligent Systems and Control, 2005.
30) Henry Anaya-Sánchez, Aurora Pons-Porrata, and Rafael Berlanga-Llavori, "A document clustering algorithm for discovering and describing topics", Pattern Recognition Letters, Vol: 31, No: 6, pp: 502-510, April 2010.
31) Congnan Luo, Yanjun Li and Soon M. Chung, "Text document clustering based on neighbors", Data & Knowledge Engineering, Vol: 68, No: 11, pp: 1271-1288, November 2009.
32) Zamir O., Etzioni O., "Web Document Clustering: A Feasibility Demonstration", in Proceedings of ACM SIGIR 98, pp. 46-54, 1998.
33) M.H.C. Law, M.A.T. Figueiredo, A.K. Jain, "Simultaneous feature selection and clustering using mixture models", IEEE Transaction on Pattern Analysis and Machine Intelligence, 26(9), pp.1154-1166, 2004.
34) Un Yong Nahm and Raymond J. Mooney, "Text mining with information extraction", ACM, pp. 218, 2004.
35) Lovins, J.B. 1968: "Development of a stemming algorithm", Mechanical Translation and Computational Linguistics, vol. 11, pp. 22-31, 1968.
36) Pant. G., Srinivasan. P and Menczer, F., "Crawling the Web". Web Dynamics: Adapting to Change in Content, Size, Topology and Use, edited by M. Levene and A. Poulovassilis, Springer- verilog, pp: 153-178, November 2004.
37) R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases", In proceedings of the international Conference on Management of Data, ACM SIGMOD, pp. 207–216, Washington, DC, May 1993.
38) R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Proceedings of 20th International Conference on Very Large Data Bases, Santiago, Chile, pp. 487–499, September 1994.
39) Bjornar Larsen and Chinatsu Aone, "Fast and Effective Text Mining Using Linear-time Document Clustering", in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, United States , pp. 16 – 22, 1999.
40) Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document Clustering Techniques", in proceedings of the KDD-2000 Workshop on Text Mining, Boston, MA, pp. 109-111, 2000.
41) B.C.M. Fung, K. Wang and M. Ester, "Hierarchical document clustering using frequent itemsets", in Proceedings of SIAM International Conference on Data Mining, 2003.
42) Reuters-21578, Text Categorization Collection, UCI KDD Archive.
43) 20-newsgroups, "http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html"
44) Webkb dataset, "http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/"

S.Murali Krishna and S.Durga Bhavani, "An Efficient Approach for Text Clustering Based on Frequent Itemsets", European Journal of Scientific Research, vol. 42, no.3, 2010 (accepted for publication).

Global Journals Inc. (US) Guidelines Handbook 2011

www.GlobalJournals.org

FELLOW OF INTERNATIONAL CONGRESS OF COMPUTER SCIENCE AND TECHNOLOGY (FICCT)

- FICCT' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'FICCT" can be added to name in the following manner e.g. **Dr. Andrew Knoll, Ph.D., FICCT, Er. Pettor Jone, M.E., FICCT**
- FICCT can submit two papers every year for publication without any charges. The paper will be sent to two peer reviewers. The paper will be published after the acceptance of peer reviewers and Editorial Board.
- Free unlimited Web-space will be allotted to 'FICCT 'along with subDomain to contribute and partake in our activities.
- A professional email address will be allotted free with unlimited email space.
- FICCT will be authorized to receive e-Journals - GJCST for the Lifetime.
- FICCT will be exempted from the registration fees of Seminar/Symposium/Conference/Workshop conducted internationally of GJCST (FREE of Charge).
- FICCT will be an Honorable Guest of any gathering hold.

ASSOCIATE OF INTERNATIONAL CONGRESS OF COMPUTER SCIENCE AND TECHNOLOGY (AICCT)

- AICCT title will be awarded to the person/institution after approval of Editor-in-Chef and Editorial Board. The title 'AICCTcan be added to name in the following manner:
  eg. **Dr. Thomas Herry, Ph.D., AICCT**
- AICCT can submit one paper every year for publication without any charges. The paper will be sent to two peer reviewers. The paper will be published after the acceptance of peer reviewers and Editorial Board.
- Free 2GB Web-space will be allotted to 'FICCT' along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted with free 1GB email space.
- AICCT will be authorized to receive e-Journal GJCST for lifetime.
- A professional email address will be allotted with free 1GB email space.
- AICHSS will be authorized to receive e-Journal GJHSS for lifetime.

ANNUAL MEMBER

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

PAPER PUBLICATION

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

Process of submission of Research Paper

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

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

The Author can submit the paper either online or offline. The authors should prefer online submission.

Online Submission: There are three ways to submit your paper:

(A) (I) Register yourself using top right corner of Home page then Login from same place twice. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal from "Research Journals" Menu.

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

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

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

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

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

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

You can use your own standard format also.

Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

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

Scope

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

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

2. ETHICAL GUIDELINES

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

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

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

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

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

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

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

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

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

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

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

Please mention proper reference and appropriate acknowledgements wherever expected.

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

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

3. SUBMISSION OF MANUSCRIPTS

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

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

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

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

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads: Original research paper: Such papers are reports of high-level significant original research work.

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

Research articles: These are handled with small investigation and applications

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

5. STRUCTURE AND FORMAT OF MANUSCRIPT

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

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

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

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

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

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

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

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

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

(h) Brief Acknowledgements.

(i) References in the proper form.

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

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

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

Format

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

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

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

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

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:
Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.
Abstract, used in Original Papers and Reviews:
Optimizing Abstract for Search Engines
Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words
A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.
One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

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

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

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

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

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

Acknowledgements: Please make these as concise as possible.

References

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

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

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

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

Tables, Figures and Figure Legends

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

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

Preparation of Electronic Figures for Publication

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

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

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

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

6. AFTER ACCEPTANCE

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

6.1 Proof Corrections

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

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

(Free of charge) from the following website:

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

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

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

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

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

6.3 Author Services

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

6.4 Author Material Archive Policy

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

6.5 Offprint and Extra Copies

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

Informal Tips for writing a Computer Science Research Paper to increase readability and citation

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

Techniques for writing a good quality Computer Science Research Paper:

1. Choosing the topic- In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish

the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be

sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

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

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

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

Informal Guidelines of Research Paper Writing

Key points to remember:

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

Final Points:

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

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

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

General style:

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

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page

- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

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

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

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

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

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

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

Title Page:

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

Abstract:

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

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

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to

shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

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

Approach:

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

Introduction:

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

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

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.
- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic

principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

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

Methods:

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

Approach:

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

What to keep away from

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

Results:

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

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

Content

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

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.

- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

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

Figures and tables

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

Discussion:

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

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

Approach:

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

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

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

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

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

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

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

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

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

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

Index

E

F

G

H

I

K

L

M

O

P

Q

R

S

T

U

V

W

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