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

Global Journal of Computer Science and Technology

DISCOVERING THOUGHTS AND INVENTING FUTURE

Technology 10 Reforming Ideas

May 2011

Pinnacles

Self-Organizing Genetic Algorithm

Web Personalization: A Review

Using Dynamic Neural Network

Up-Down Routing Based Deadlock

The Volume 11
Issue 7
Version 1.0

ENG

Global Journal of Computer Science and Technology

Global Journal of Computer Science and Technology

Volume 11 Issue 7 (Ver. 1.0)

Global Association of Research

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

Global Journals Inc.

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

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

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

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

eContacts

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

Pricing (Including by Air Parcel Charges):

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

Editorial Board Members (HON.)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Contents of the Volume

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

Neural Web Based Human Recognition

By M. Prabakaran, Dr. T. Senthil Kumar
Vinayaka Missions University

Abstract- Face detection is one of the challenging problems in the image processing. A novel face detection system is presented in this paper. The approach relies on skin-based color features xtracted from two dimensional Discrete Cosine Transfer (DCT) and neural networks, which can be used to detect faces by using skin color from DCT coefficient of Cb and Cr feature vectors. This system contains the skin color which is the main feature of faces for detection, and then the skin face candidate is examined by using the neural networks, which learn from the feature of faces to classify whether the original image includes a face or not. The processing is based on normalization and Discrete Cosin Transfer. Finally the classification based on neural networks approach. The experiment results on upright frontal color face images from the internet show an excellent detection rate.

Keywords: *Face detection, skin color segmentation, compressed domain , neural networks.*

GJCST Classification: *I.4.6, I.5.1*

Strictly as per the compliance and regulations of:

Neural Web Based Human Recognition

M. Prabakaran[α], Dr. T. Senthil Kumar[Ω]

Abstract- Face detection is one of the challenging problems in the image processing. A novel face detection system is presented in this paper. The approach relies on skin-based color features extracted from two dimensional Discrete Cosine Transfer (DCT) and neural networks, which can be used to detect faces by using skin color from DCT coefficient of Cb and Cr feature vectors. This system contains the skin color which is the main feature of faces for detection, and then the skin face candidate is examined by using the neural networks, which learn from the feature of faces to classify whether the original image includes a face or not. The processing is based on normalization and Discrete Cosin Transfer. Finally the classification based on neural networks approach. The experiment results on upright frontal color face images from the internet show an excellent detection rate.

IndexTerms - *face detection, skin color segmentation, compressed domain , neural networks.*

I. Introduction

Face detection is an active area of research spanning disciplines such image processing,, pattern recognition and computer vision .face detection and recognition are preliminary steps to wide of applications such as personal identity, video surveillance etc. the detection efficiency influences the performance of these systems, there have been various approaches for face

Detection, which classified into four categories

(i) knowledge based method
(ii) feature based method
(iii) Template matching method
(iv) Appearance based method .a comprehensive survey of the face detection given here.

In the compressed domain chrominance, shape and DCT information coefficient was combined by Wang and Chang to achieve high-speed face detection without decoding of the compressed video image. The proposed technique derived from [1], in their works a direct access content and extraction features in compressed domain instead of pixel domain. The algorithm works directly on the DCT coefficient parameters, DCT coefficient as features based compression reduce spatial redundancy and captures the compact information about the patterns .color information is used as the main detection clues, a skin color model is created in the level of y cb cr color space .The reason for choosing Cb and Cr Color space that there is no information about luminance, classification using only pixel chrominance, skin segmentation may become more robust to lighting variations if pixel luminance is discarded and speed up the calculation in detecting the skin face regions.

The objectives of this research are to develop better normalization method and also aim to improve the segmentation that will assist and quick detecting faces from images. And also to implement a classifier face based on neural networks for face detection. Most of the interest reader are referred to the comprehensive survey on face detection by Yang et al,and by Hjelmas and Low .The new algorithms introduced combines two methods to perform fast and accurate face detection system , which are a feature based methods and image based methods, the feature based method used a pre-processor of the image based method and guides the search of image based methods using neural networks that examine the face candidate regions instead of performing huge search in every part of the test image.Hwei proposed Extraction regions of skin can be either pixel–based or region based .The diagram of our proposed techniques is presented in fig .1 skin segmentation is applied using the predefined color range threshold of Cb and Cr range .2D Discreet Cosine Transfer (DCT) for each sub-block image is computed and features vector are formed from the DCT coefficients ; where DCT can be as signature useful for recognitions tasks such as facial expression recognitions.

II. Face Detection in Image

Many techniques for face detection in image were classified into four categories Knowledge based method It dependence on using the rules about human facial feature .It is easy to come up with simple rules to describe the features of a face and their relationships. For example, a face often appears in an image with two eyes that are symmetric to each other, a nose, and a mouth. , and features relative distance and position represent relationships between feature. After detecting features, verification is done to reduce false detection. This approach is good for frontal image; the difficulty of it is how to translate human knowledge into Known rules and to detect faces in different poses.

a) *Image Based method*

In this approach, there is a predefined standard face pattern is used to match with the segments in the

About[α]- Research Scholar,Vinayaka Missions University, Salem.
E-mail- captainprabakaran@gmail.com
About[Ω]- Reader, Head, Dept of Automobile Engineering Anna University-Tiruchirappali, Tamil Nadu,
E-mail- 1tsenthil@gmail.com

image to determine whether they are faces or not. It uses training algorithms to classify regions into face or non-face classes. Image-based techniques depends on multi-resolution window scanning to detect faces, so these techniques have high detection rates but slower than the feature-based techniques. Eigen-faces and neural networks are examples of image-based techniques. This approach has advantage of being simple to implement, but it cannot effectively deal with variation in scale, pose and shape

b) *Features Based method*

This approach depends on extraction of facial features that are not affected by variations in lighting conditions, pose, and other factors. These methods are classified according to the extracted features [1]. Feature-based techniques depend on feature derivation and analysis to gain the required knowledge about faces. Features may be skin color, face shape, or facial features like eyes, nose, etc.... Feature based methods are preferred for real time systems where the multi-resolution window scanning used by image based methods are not applicable. Human skin color is an effective feature used to detect faces, although different people have different skin color, several studies have shown that the basic difference based on their intensity rather than their chrominance. Textures of human faces have a special texture that can be used to separate them from different objects. Facial Features method depends on detecting features of the face. Some users use the edges to detect the features of the face, and then grouping the edges. Some others use the blobs and the streaks instead of edges. For example, the face model consists of two dark blobs and three light blobs to represent eyes, cheekbones, and nose. The model uses streaks to represent the outlines of the faces like, eyebrows, and lips .Multiple Features methods use several combined facial features to locate or detect faces. First find the face by using features like skin color, size and shape and then verifying these candidates using detailed features Such as eye brows, nose, and hair.

c) *Template matching method*

Template matching methods use the correlation between pattern in the input image and stored standard patterns of a whole face / face features to determine the presence of a face or face features. Predefined templates as well as deformable templates can be used.

III. Face Detection Algorithms

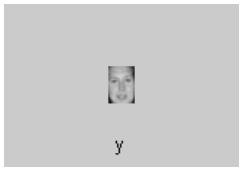
Information of skin color in a color image is a very popular and useful technique for face detection. The obvious advantage of this method is simplicity of skin detection rules that leads to construction of a very rapid classifier. We can use color information as a feature to identify a person's face in an image because human faces have a special color distribution that differs significantly, although not entirely, from those of the background objects. Previous studies have found that pixels belonging to skin region exhibit similar chrominance components within and across different human races. In the YCbCr color space, chrominance components are represented by Cb and Cr values. Thus, skin color model can be derived from these values. By using threshold techniques , skin color pixels are identified by the presence of a certain set of Cb and Cr values which corresponding to the respective ranges of RCb and RCr values of skin color. Otherwise, the pixel is classified as non skin color. The system being designed into three main categories, preprocessing , segmentation, classification using neural Networks.

a) *Pre-Processing*

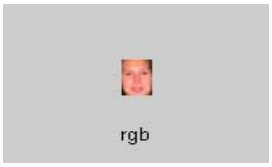
In fact, processing skin color is faster than other facial features, collecting a data set of skin face by cropping or cutting manually the image skin face and non-skin face to get a dataset of face and non-face. Different people have different skin color, while the difference lies mostly in the color intensity not in chrominance color itself. Literature survey show that Y Cb Cr color space is one of the successful color spaces in segmenting skin color accurately .Selecting the suitable color space to model skin color and a void variation of lighting condition Cb and Cr Color space. Extract DCT coefficient features from Cb and Cr blocks

b) *Segmentation Skin Color*

Skin color information is very important features for many researches, however the accuracy of skin color detection is important for face detection [2]. In this paper we convert the image from RGB to ycbcr .where are RGB is sensitive to the variation of intensity. Many skin detection method ignore the luminance component of the color space, to achieve independent model of the differences in skin appearance that may arise from the difference of human race, and also reduce the space dimension. After collecting a different human faces and analyzing the histogram distribution sample skin color values of chrominance component to represent the likelihood of the pixel belonging to the skin region.it was found that the chrominance component of the skin color fails in a certain range .X is skin color [1], if its projection on the Cb and Cr plane is inside predetermined rectangle $Cb \in Rcb$ and $Cr \in Rcr$ i.e., $r1\ r\ r2\ C \leq C \leq C$ and $b1\ b\ b2\ C \leq C \leq C$ .where $cb\ R = [\ b1\ C\ ,\ b2\ C\ ]$ and $R_{cr} = [\ C_{r1}\ ,\ C_{r\ 2}\ ]$,which are found experimentally used to eliminate quickly non-skin face color. And also to improve the segmentation of skin color regions Fig.2 shows the distribution histogram skin regions sample and the threshold of for Cb and Cr color space.



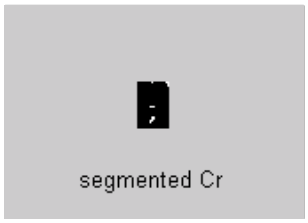
(a) RGB image



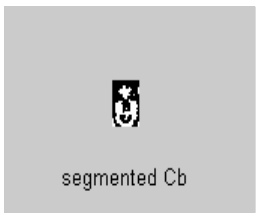
(b) Y component

(c) Cr component

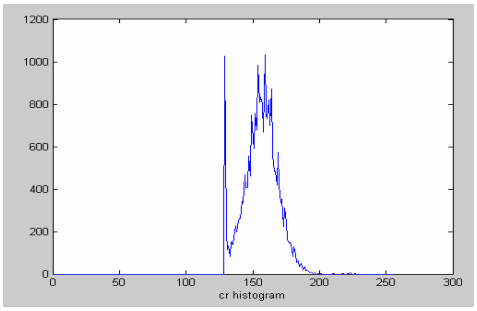
(d) Cb component



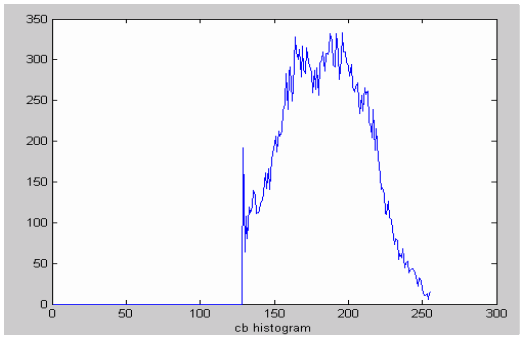
(e)region segmented on Cb



(f)region segmented on Cr



(g) Cr histogram distribution sample



(h) Cb histogramdistribution sample

Fig.2 skin face region segmentation

c) *Feature Extraction*

Discrete cosine transform is used widely in many applications and mainly used in the compressed data domain. and forms the basis well known JPEG image compression format. Jiang el.al [1] introduced simple low cost and fast algorithms that extract dominant color feature directly fro DCT rather than in the pixel domain the extracted DCT Coefficient can be used as type of a signature of which might be useful for recognition task, such as facial expression recognition [5]. The proposed technique derived from [1], The system calculates the 2D-DCT for each cropped skin block coming out of the previous stage. This results in a matrix of 1 × 48 coefficients of both Cb and Cr color space components within the processed image block.,. Which are these values is taken to construct the feature vector. Empirically, the upper left corner of the 2D-DCT matrix contains the most important values, because they correspond to low-frequency, however the upper most coefficient is called DC and it correspond to average light intensity of the block. The others are called AC, and those coefficient provide useful information about the texture detail in the blocks. For each block we use the DC's and the first three zig zag order AC's as a set of 1 × 4 vector coefficients as shown in fig.4.

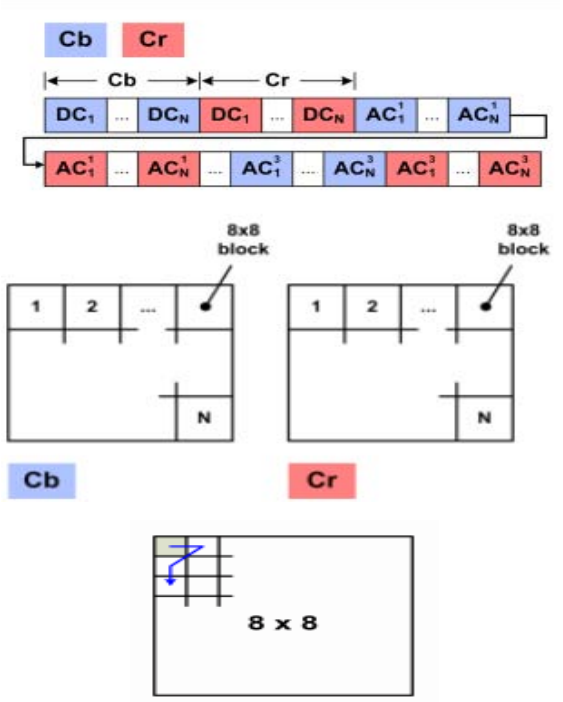


Fig.4. feature extraction from DCT coefficient

d) *Classification*

Neural networks are often used in face detection, Rowley, Baluja, Kanade [4] proposed a face detection methods based on neural networks that could discriminate between face and non face on large dataset images.

In our system, we use (MLP) multi layer perception back propagation neural networks in order to training data set and classify features that are extracted using DCT(Discreet Cosine Transfer coefficient). After divided into blocks of size 8x8 pixels. Training using a vector obtained from 18X27 training data set of 8x8 pixel block for true oval face may usually guarantees that only pixels the face are used as input to neural networks, however, to produce an

output of 0.9 for the skin face and 0.1 for the non-skin face after repeatedly presented input samples and desired targets, compared the output with the desired and measuring the error and adjusting the weights until correct output for every input[4].The main advantage of choosing Backprobagation neural networks the simplicity and capability in supervised pattern matching.

IV. Neural Networks

Neural networks have been applied in many pattern recognition problems like object recognition .there is many image based face detection using neural networks the most successful system was introduced by Rowley et al [4] as using skin color segmentation to test an image and classify each DCT based feature vector for the presence of either a face or non face .

The neural networks used in this paper back propagation neural networks and was chosen because of simplicity and its capability in supervised pattern matching. The structure of the neural network with three layers, the input layer is a vector of 1xn DCT coefficient vectors of neuron from each image either face or non face image, the hidden layers has n neurons, and the output layer is a single neuron which is 0.9 if the face is presented and 0.1 otherwise. The neural networks is trained using DCT coefficient feature vectors after skin face color candidate obtained from the segmentation stage, which are the DC and the first three zig zag order AC's features samples from each blocks 8x8 pixels of an manually cropped image 18x27 pixel of face and non- faces to classify each feature vector as output value 0.9 for a face and 0.1 for non-face.

V. Expriments and Results

We show in the section a set of experimental results to presents the performance of the proposed system, the experimented was implemented using Mat lab Version 7.2 on the Intel Pentium(4) 2.80Ghz 1.00GB of RAM and Windows XP operating system .This section presents results of experiment applied on the unknown input test image containing a face or non-face. Starting with sliding overlapping window 18x27, by overlap scanning the window, where different overlap parameter used 1,2 up to the half pixels, in our experiment 9 pixel is the half of the window it might be maximum overlap, then each part of the unknown test image is scanned using slid window and extracted the DCT features and feed it to the trained neural networks of the dataset of images. However the neural networks tested with the trained neural networks and classify it to see if the part containing a face or non face. The experiment results shows that our face detection system is reliable that neural networks able to detect and classify pattern features accurately under different overlap sliding scan window over the unknown input test image. The converge response of training dataset shows accurate and excellent face and non classification as in fig.4a , In fig.4b the result is reasonable, since the test set error and the validation set error have similar characteristics , and The next step is to perform some analysis of the network response. By putting the entire data set through the network (training, validation and test) to perform a linear regression between the network outputs and the corresponding targets as shown in fig.4c, according to the excellent response of the Backprobagation neural networks with the target desired ,the classification performance provides a comprehensive excellent picture of the classification performance of the classifier.

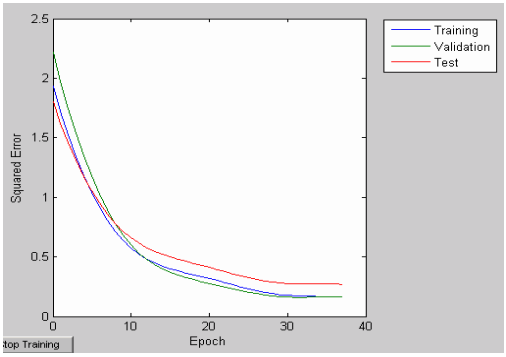


Fig 5. NN's trained with the training, cross-validation and testing

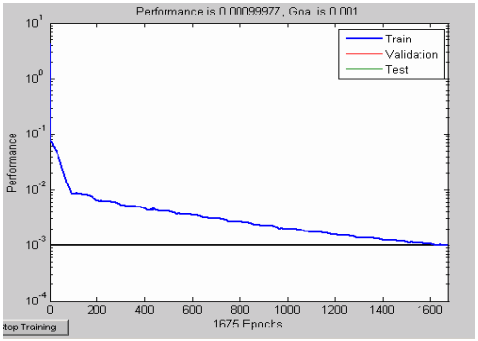


Fig 6. Training performance

VI. Conclusions and Future Work

This paper proposes a new algorithm for face detection in the compressed domain , extracted DCT coefficient vector features after segmentation a face skin candidate using skin color information on both Cb Cr color space, along with backprobgation neural networks classifier. We have divided the problem into three stages pre-processing, segmentation, and classification using back propagation neural networks.

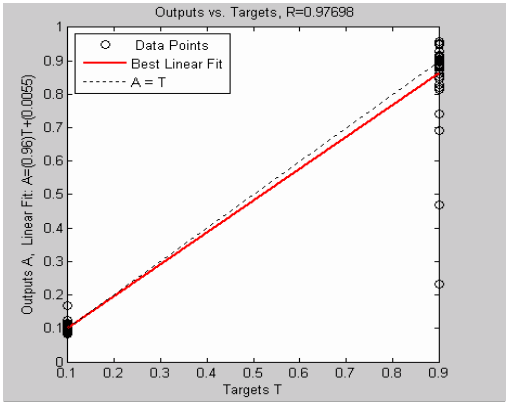


Fig 7. Linear Regression

The system has been tested on a dataset of upright frontal color face images from the internet and achieved excellent detection rate. These methods as a future work, will improve the detection of faces in compressed images to be use for face image retrieval based on skin color and also we may split the features DC's and AC's and feed it as two inputs to the neural networks. However the system proposed can be used as first step to face recognition system.

References Références Referencias

1. Jiamin Jiang,Ying Weng,and P.LI " Dominant colour extraction in DCT Domain" in Image and Vision computing 24 , 2002 ublished by Elsevier B.V,pp.129-177.
2. Hwei-Jen Lin, Shu-Yi Wang, Shwu-Huey, and Yang –Ta-Kao "Face Detection Based on Skin Color Segmentation and Neural Network" *IEEETransactions on*, Volume: 2, pp544- 549, ISBN: 0-7804-9422-4
3. Y. Ming-Hsuan, D. J. Kriegman, and N. Ahuja, "Detecting faces in Images: a survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 44-58, 2002.
4. H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," in *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, 1998, pp. 24-48.
5. L. Ma, Y. Xiao, K. Khorasani, and R. K. Ward, "A new facial expression recognition technique using 2D DCT and k-means algorithm", in *Proc.International Conference on Image Processing*,Oct. 2004, pp. 129-172.
6. F.Smach, M.Atri, J.Miteran and M.Abid "Design of a Neural Networks Classifier for Face Detection" in Journal of computer Science 2(3):pp257-260, 2006
7. Lamiaa Mostafa, Sharif Abdelazeem "Face Detection Based on Skin Color Using Neural Networks" in GVIP 05 Conference, pp19-21, Dec 2006, CICC, Cairo, Egypt.
8. V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques", in *Proc. Graphicon-2003*
9. H. Kruppa, M. A. Bauer, and B. Schiele, "Skin patch detection in Realworld images", in *Proc. of the DAGM-Symposium*, 2002, pp. 109-116.
10. L. Ma, Y. Xiao, K. Khorasani, and R. K. Ward, "A new facial expression recognition technique using 2D DCT and k-means algorithm", in *Proc.International Conference on Image Processing*, Oct. 2004, pp. 1269- 1272.
11. L. Ma and K. Khorasani, "Facial expression recognition usingconstructive feedforward neural networks", *IEEE Transactions on Systems,Man and Cybernetics*, Part B, Vol. 34, No. 3, June 2004, pp.1588 – 1595
12. L. Ma, Y. Xiao, K. Khorasani, and R. K. Ward, "A new facial expression recognition technique using 2D DCT and k-means algorithm", in *Proc.International Conference on Image Processing*,Oct. 2004, pp. 1269-1272.
13. Jiamin Jiang,Ying Weng,and P.LI " Dominant colour extraction in DCT Domain" in Image and Vision computing 24 , 2006 published by Elsevier B.V,pp.1269-1277.
14. E. Hjelmås, and B. K. Low, "Face detection: a survey", *Computer Vision and Image Understanding,* Vol. 83, No. 3, Sept. 2001, pp. 236-274.
15. H. Wang and S. –F. Chang, "A highly efficient system for automatic face region detection in mpeg video," In *IEEE Trans*. CSVT, 7(4), 1997.

May 2011

6

Global Journal of Computer Science and Technology Volume XI Issue VII Version I

This page is intentionally left blank

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

Self-Organizing Genetic Algorithm for Multiple Sequence Alignment

By Amouda Nizam, Buvaneswari Shanmugham, Kuppuswami Subburaya
Pondicherry University

Abstract- Genetic algorithm (GA) used to solve the optimization problem is self-organized and applied to Multiple Sequence Alignment (MSA), an essential process in molecular sequence analysis. This paper presents the first attempt in applying Self-Organizing Genetic Algorithm for MSA. Self-organizing genetic algorithm (SOGA) can be developed with the complete knowledge about the problem and its parameters. In SOGA, values of various parameters are decided based on the problem and fitness value obtained in each generation. The proposed algorithm undergoes a self-organizing crossover operation by selecting an appropriate rate or a point and a self-organizing cyclic mutation for the required number of generations. The advantages of the proposed algorithm are (i) reduce the time requirement for optimizing the parameter values (ii) prevent execution with default values (iii) avoid premature convergence by the cyclic mutation operation. To validate the efficiency, SOGA is applied to MSA, and the resulting alignment is evaluated using the column score (CS). The comparison result shows that the alignment produced by SOGA is better than the widely used tools like Dialign and Multalin. It is also evident that the proposed algorithm can produce optimal or closer-to-optimal alignment compared to tools like ClustalW, Mafft, Dialign and Multalin.

Keywords: *Crossover, Genetic Algorithm, Multiple Seuence Alignment, Mutation, Selection, Self organization*

GJCST Classification: *J.3*

Strictly as per the compliance and regulations of:

Self-Organizing Genetic Algorithm for Multiple Sequence Alignment

Amouda Nizam[α], Buvaneswari Shanmugham[Ω], Kuppuswami Subburaya[β]

Abstract- Genetic algorithm (GA) used to solve the optimization problem is self-organized and applied to Multiple Sequence Alignment (MSA), an essential process in molecular sequence analysis. This paper presents the first attempt in applying Self-Organizing Genetic Algorithm for MSA. Self-organizing genetic algorithm (SOGA) can be developed with the complete knowledge about the problem and its parameters. In SOGA, values of various parameters are decided based on the problem and fitness value obtained in each generation. The proposed algorithm undergoes a self-organizing crossover operation by selecting an appropriate rate or a point and a self-organizing cyclic mutation for the required number of generations. The advantages of the proposed algorithm are (i) reduce the time requirement for optimizing the parameter values (ii) prevent execution with default values (iii) avoid premature convergence by the cyclic mutation operation. To validate the efficiency, SOGA is applied to MSA, and the resulting alignment is evaluated using the column score (CS). The comparison result shows that the alignment produced by SOGA is better than the widely used tools like Dialign and Multalin. It is also evident that the proposed algorithm can produce optimal or closer-to-optimal alignment compared to tools like ClustalW, Mafft, Dialign and Multalin.

Keywords- Crossover, Genetic Algorithm, Multiple Sequence Alignment, Mutation, Selection, Self-organization.

I. Introduction

Self-organizing system functions without any guidance from the external control (without a central control). Self-organization is done based on local information obtained from the interactions of lower-level components [1]. It is evident from the literature, several GA, a stochastic iterative method [2] are proposed for MSA, to align set of sequences. Major problem of GA, premature convergence can be avoided by blending the concept of self organization and GA. Using SOGA several other problems are solved but applying for MSA with a new mechanism is first of its kind.

About[α]- Centre of Excellence in Bioinformatics, School of Life Sciences, Pondicherry University, Puducherry – 605 014 (corresponding author phone: +91-413-2655212; fax: +91-413-2655211;
E-mail: amouda@yahoo.com
About[Ω]- Centre of Excellence in Bioinformatics, School of Life Sciences, Pondicherry University, Puducherry
E-mail: buvanisuriya@bicpu.edu.in).
About[β]- Department of Computer Science, School of Mathematics and Computer Science, Pondicherry University, Puducherry
E-mail: skswami@yahoo.com

In this case, MSA is defined by the position and gap size in the sequences. Two types of search operators like recombination and gap mutation are included in the algorithm to produce offspring alignments [3]. Apart from these two basic operators, several operators are also proposed in the literature to improve the performance of GA [4-5]. In some case existing GA operators are unsuitable as they are not specific for the problem and the encoded chromosome. This led us to develop new GA operators, specifically for MSA.

The proposed algorithm is illustrated using DNA sequences, but it can be extended to RNA and protein sequences also. A set of *n* DNA sequences of varying length are considered for the alignment process. The nucleotide bases *A*, G, *C*, T corresponds to adenine, guanine, cytosine and thymine and gaps are represented by '-' (hyphen).

The remainder of the paper is organized as follows. The next two section reviews multiple sequence alignment and genetic algorithm. Section 3 explains various methods of the self-organizing genetic algorithm and its advantages over standard GA. Section 4 explain the proposed SOGA with its pseudocode. Section 5 explains the working mechanism of SOGA-MSA with newly developed operators. Section 6 shows the comparison results and discussion. Section 7 is the conclusion and future perspectives.

II. Multiple Sequence Alignment

MSA, aligning three or more nucleotide or amino acid sequences simultaneously is one of the important tasks in bioinformatics. Important application of MSAs is their incorporation in many structure and function prediction methods from sequence. It can reveal conserved residues that enable the identification of possibly important sites. The construction of MSA is closely related to phylogenetic analysis and a phylogenetic tree can be inferred by MSA. The study of molecular evolution is an area where MSA is extensively used [6].

The computation of an optimal alignment mathematically is too complex. Current implementation methods are heuristics in which full optimization is not guaranteed. Various algorithms available for MSA are classified into three main categories: Exact, Progressive and Iterative based on their properties.

Exact algorithms are high quality heuristic in nature, produce very close to optimal alignment. It can handle the only restricted number of sequences and are limited to sums-of-pairs as an objective function.

Progressive alignment using dynamic programming depends on a progressive assembly of the multiple alignments, heuristic in nature but does not guarantee any level of optimization.

Iterative alignment methods produce alignment and refine it through a series of cycles (iterations) until no further improvements can be made. It is deterministic or stochastic depending on the strategy used to improve the alignment. It allows for a good conceptual separation between optimization processes and objective function as its main advantages[7].

The widely used MSA tools implementing different algorithms are ClustalW[8], MultAlin[9], DIALIGN[10], MUSCLE[11], T-Coffee[12], DCA[13]. In addition GA based MSA software like SAGA[14], MSA-GA[3] are available but not in an executable form.

(i) Its flexibility in assigning the fitness function, mathematical function used to evaluate the fitness of the chromosomes.
(ii) The complexity of the MSA process increase exponentially, NP-hard (nondeterministic polynomial) in nature[7] can be solved by GA.
(iii) It is not restricted to need of a particular algorithm to solve the problems. Needs only fitness function to evaluate the chromosomes [15].

III. Genetic Algorithm

GA starts with the generation of population consists of chromosomes, a fixed size encoded solution. Each chromosome represents a possible solution and the space of all feasible solutions is called search space. The role of GA is to alter the generated chromosomes using various operators to get the optimal chromosome with best fitness value in the search space. Iteration continues till the termination condition is satisfied.

Outline of basic GA

1. **[Start]** Generate random population of *n* chromosomes.
2. **[Fitness]** Evaluate fitness *f(x)* of each chromosome *x* in the population.
3. **[New Population]** Create new population using (i to iv) repeatedly until the process is complete
   i) **Selection**
   ii) **Crossover**
   iii) **Mutation**
   iv) **[Accepting]** Place new offspring in a new population.
4. **[Replace]** Use newly generated population for the next iteration.
5. **[Test]** Check the termination condition, if satisfied, stop, and return the best solution.
6. **[Loop]** Go to step 2[16].

IV. Self-Organizing Genetic Algorithm (SOGA)

For a specific input, setting the GA parameters is an important task. The concept of self-organizing GA is to adapt values for parameters like population size, number of generations, selection modes, rates of selection crossover and mutation during execution.

In the blend of SO and GA, most of the parameters change according to the fitness of the chromosomes. An attempt towards SOGA requires a complete understanding of the relationship among various parameters and its impact in the performance.

The aim of SOGA is to create an automated computer program that solves the problem with little or no information from the user. The difficulty in choosing the appropriate number of generations, chromosome length, crossover and mutation rate is eliminated, thus GA is made efficient and simple to use.

Using GA, solutions to a particular problem are not algebraically calculated rather found by a population of solution alternatives, which are altered (using operators like crossover and mutation) in iterations of the algorithm in order to increase the probability of having better solutions. In optimization, better chromosomes with higher fitness value will be selected.

SOGA over Standard Genetic Algorithm (SGA)

Encoding of chromosomes in SGA is usually fixed-length strings. In SOGA, the length of the chromosome can be made to change adaptively based on the problem[17].

Population Size is fixed in SGA and the corresponding number of chromosomes is generated. Population size 50-100 is reported as best[18]. Population size can be made to change adaptively based on the problem.

- It can be self- organized by generating both small and large populations and the fitness value of each of the chromosomes is calculated. If the average fitness of the larger population is higher than the smaller population then the program continues with the larger population otherwise with the smaller population.
- Each time at convergence, population size is doubled till it reaches an upper limit[19].

Number of Generations is always fixed in SGA, and the algorithm terminates on reaching specified number of generations or fitness level or at convergence. An optimal solution may not be reached if termination is due to the maximum number of generations. Hence it

is necessary to self-organize number of generations based on the problem.

Selection Operator in SGA is usually one or combination of operators. In SOGA, certain conditions are defined to choose the appropriate operator for a particular problem for e.g. based on the average fitness of the generated chromosome.

Crossover/ Mutation Operator in SGA is usually one or combination of operators, and it can be self-organized

- By defining conditions based on which the appropriate operator or rate is chosen.
- Crossover/ Mutation operation is performed with a specified number of methods and based on the average fitness of the resulting chromosome, an appropriate method is chosen[20].
- The algorithm can be executed initially with a minimum optimal crossover/ mutation rate. At each point of convergence, instead of termination the rate can be increased cyclically till it reaches the optimal upper limit[21].
- The crossover/ mutation rates adapted from high to minimum optimal rate [22].
- Along with the chromosome generated with the current value obtained by increase or decrease in the rate, chromosomes corresponding to larger and smaller are also generated. The chromosome with higher fitness is chosen[23] as an elite.

It is reported in the literature that generally crossover rates should be high (80%-90%) and mutation rate should be very low (0.5%-1%)[18].

Advantage of SOGA

- GA with self-organizing coding, operators and parameter values is efficient and simple to use.
- Time required for optimizing parameter values is eliminated by using SOGA. In SGA, optimal parameter value can be found by executing with all possible values and combinations with other parameters.
- The default parameter values assumed to be optimal is considered when the user fails to select appropriate values. Even this value may lead to bad results for some problems. Instead of getting parameter values from user SOGA self-organizes all or most of the parameters to assign values based on the problem.

V. Proposed SOGA

SOGA proposes two new operators to perform crossover and mutation. The proposed Self-Organizing Crossover Operator **(SOCO Operator)** selects an appropriate crossover point and the corresponding rate from the initial crossover point. The proposed mutation operator **(Self-Organizing Binary Shuffler)** converts the chromosome representation into a binary form and performs mutation for a range of rates till the termination condition is satisfied. The number of generations is also self-organized, which varies depending on the problem.

Pseudo code of the SOGA

1. **[Start]** Generate the random population of *n* chromosomes.
2. **[Fitness]** Evaluate the fitness *f(x)* of each chromosome *x* in the population.
3. **[Selection]** Select and save the elite (chromosome with highest fitness value) in the current population.
4. **[New Population]** Create a new population using (i to iv) repeatedly until the process is complete

i) **[SOCO]** Self-organizing the selection of crossover point based on the specified optimal rate and perform crossover using **SOCO operator**.

ii) **[Selection]** Select and save the elite in the current population.

iii) **[SOBS]** Convert the chromosome representation to a binary form and perform mutation for a range of rates cyclically using **SO Binary Shuffler**.

iv) **[Selection]** Select and save the elite in the current population.

5. **[Test]** Check the termination condition. If satisfied, stop, and return the best solution.
6. **[Loop]** Go to step 4.

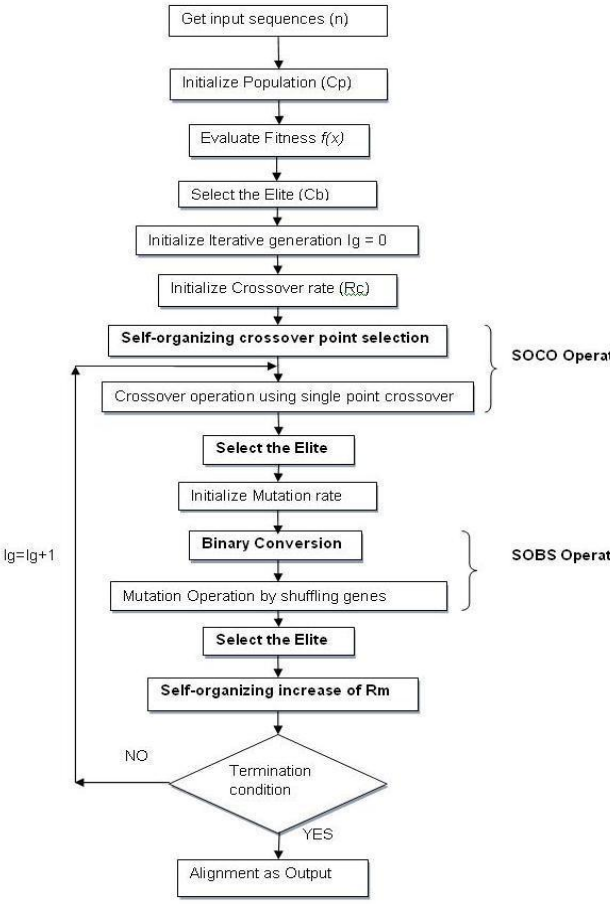


Fig.1. Flow chart of SOGA.

VI. SOGA-MSA

a) *Chromosome Representation*

In general, chromosome is a matrix with fixed lengths and represented as sequences with spaces[25,26]. For the problem of MSA, the gap positions are used to encode the chromosome. The number of gaps to be inserted in each sequence is calculated in such a way that the length of all sequences in the alignment (global) is same. A single chromosome consists of gap positions of all the sequences in order. In the mutation process, the chromosomes are encoded as binary digits (1, 0) representing presence and absence of the gap in sequence. In SOGA, the length of the chromosome is adaptively changed based on the number of sequences and its length[17].

b) *Number of Generations*

In each generation, the algorithm generates chromosomes of required population, and its fitness score is evaluated. Chromosomes from the current population are stochastically selected and modified by crossover and mutation, which undergo next generation. As the rate of mutation is made to increase cyclically based on the fitness value, iteration completes only when the optimal upper limit is reached. The number of generations depends on the betterment of the fitness value obtained in each generation.

For e.g., consider the dataset 469 with three sequences as input. The generation starts with Rm=1% produces an alignment with CS = 36. Next generation continues with Rm 1% resulting further no increase in CS. Hence by the concept of self-organization Rm increased to 3% resulting CS = 37. Self-organizing process continues till the upper limit of Rm (80%) is reached. In 43 generations the CS of the output alignment is 45 as shown in the table I.

Table 1: Example For Self-Organizing Number Of Generations

Iterative Generation (Ig)	Mutation rate (Rm)	Column Score (CS)
1	1%	36
3	3%	37
21	37%	40
25	43%	44
41	79%	45
43	81%	45

c) *Population Initialization*

Population size indicates the number of chromosomes in a generation, and it must be optimal for a particular problem.

Table 2: Example For Population Initialization

Sequence	Sequence Length	No. of Gaps	Gap positions	Sorted gap positions	Alignment
TCTAGATG	8	6	5|0|11|3|6|9|	0|3|5|6|9|11|	-TC-T--AG-A-TG
CTATGATGTA	10	4	12|10|0|7|	0|7|10|12|	-CTATGA-TG-T-A
ACGATGTA	8	6	7|4|11|5|8|13|	4|5|7|8|11|13|	ACGA--T--GT-A-
GTTCTAT	7	7	8|4|6|1|13|3|0|	0|1|3|4|6|8|12|	--G--T-T-CTA-T
ACGTATAGCAAT	12	2	9|4|	4|9|	ACGT-ATAG-CAAT

Best population size also depends on encoding method and size of the encoded string. According to research, it is proven that increase in population size after a limit does not improve the performance of GA[18].

Considering m sequences to be aligned with the length ($m_1,\ldots m_i$) and the space ratio r_{sp} =0.2. If the longest length of sequences to be aligned is m_{max}, then N = m_{max} * (1+ r_{sp}). The value of N is the size of search space of alignments. It limits the longest length of alignments that chromosomes can represent.

Chromosomes can be transformed to actual alignments by inserting gaps in the appropriate positions. For e.g., m_{max} = 12, r_{sp} = 0.2, then N = 14.

d) *Fitness Evaluation*

The fitness function returns a numerical score indicating fitness of the candidate alignment. It is an important parameter to determine which alignment will survive in the next generation. The fitness is evaluated by calculating the (CS) column score. CS = EM/ AL, where AL is the alignment length, EM (Exact match) =1, when all the base pair in the entire column is aligned with the same base pair.

e) *Selection*

SOGA implements an elitism operator, where an elite is the chromosome with best fitness value. The process comprises the following processes

(i) Evaluate the column score.
(ii) Sort the chromosomes.
(iii) Select and save the elite.

With the current population, SOGA undergoes a crossover. New chromosomes are generated and the elite is selected. If the fitness value of new one is greater, elite is replaced else the process continues with the same. In the same way for mutation elite selection and comparison process is repeated. This process continues for every generation to ensure that the elite saved at the end is best.

A new mechanism is followed for crossover and mutation operation in self organizing GA.

f) *Self-Organizing Crossover Operator (SOCO)*

In single point crossover operation [22], the crossover point is selected initially for a particular rate. Then the genes from starting point to the crossover point are copied from one chromosome and the rest from the second chromosome.

In SGA, crossover for a particular rate may lead to the occurrence of crossover point within a sequence itself. It may create problems like

(i) Increase in the number of gaps for a particular sequence.
(ii) Occurrence of repeated gap positions in a sequence.

To overcome this major disadvantage, proposed operator SOCO defines a new point called complete point. Each complete point refers to the end position of each sequence in a chromosome. The number of complete point in a chromosome is based on the number of input sequences. For e.g. if input sequences are five, then the chromosome contains four complete points as shown in Fig.2.

The new working principle followed by SOCO operator is as follows:

(i) Initialize the crossover rate (Rc) and select the corresponding crossover point.
(ii) Selects the complete point near the default crossover point.
(iii) Performs single point crossover operation.
(iv) Generates MSA corresponding to the chromosome.
(v) Calculates fitness score.
(vi) Selects elite.

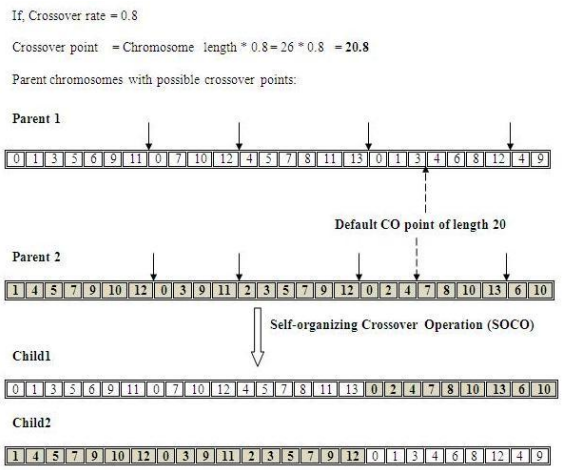


Fig.2. Example for Self-Organizing Crossover.

g) *Self-Organizing Binary Shuffler (SOBS)*

In SGA, either an optimal mutation rate which is unsuitable for all inputs is fixed or selected from a range of rates given as optional. It is hard for the user to select appropriate rate without the knowledge of the problem. To eliminate these problems, a new mutation operator with a different approach is proposed. Instead of a fixed rate, the operator performs mutation for a range of rates cyclically [21, 24] till the termination condition are satisfied.

In default shuffling process for mutation leads to the problem like

(i) Increase in the number of gaps for a particular sequence.
(ii) Occurrence of repeated gap positions in a sequence.

To avoid this, proposed mutation operator involves conversion of chromosome representation to binary digits (1,0) represents the presence and absence of gaps. The new working principle followed by SOBS operator is as follows:

i) Converts the chromosome representation to a binary form.
ii) Initialize minimum optimal mutation rate

and the corresponding mutation point is selected.

iii) Genes before mutation point are considered for mutation.

iv) The genes within each complete point and if any gene occurs between the last complete point and mutation point are shuffled separately as shown in Fig. 3.

v) Change chromosome representation to gap positions.

vi) Generates MSA corresponding to the chromosome.

vii) Calculates fitness score.

viii) Selects elite.

If the elite is replaced by the selection condition, the generation continues with the same rate else increases cyclically until an optimal upper limit is reached. The algorithm terminates on reaching the optimal upper limit when no further increase in the column scores.

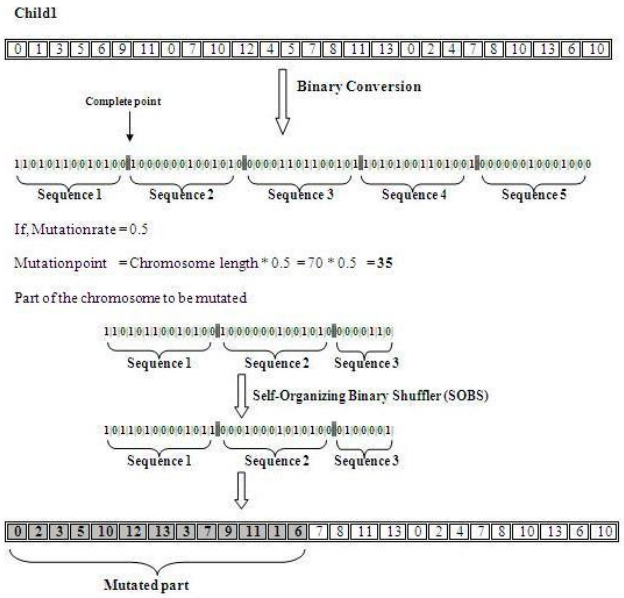


Fig.3. Example for Self-Organizing Mutation

VII. Results and Discussion

To validate the proposed algorithm, various parameters and results of SOGA is compared with SGA. The dataset 469 from oxbench_mdsa_all with three sequences is used as input. The comparative results show that on average, SOGA-MSA produces better results than the SGA-MSA. As an advantage, in fewer numbers of generations and time SOGA-MSA produces the better alignments as tabulated below.

Table 3: Comparison of Sga and Soga on Msa

GA	Population size	Crossover rate (Rc)	Mutation rate (Rm)	No. of Generations	Exact match (EM)	Alignment Length (AL)	Column Score (CS)
SGA-MSA	100	80	60	50	44	406	0.10
SOGA-MSA	100	70	1-80	43	45	406	0.11

The alignments produced by the widely used MSA tools with default parameter settings are compared with the developed **SOGA-MSA,** and the results are tabulated. The standard reference datasets of DNA sequence alignments from BAliBASE[27] are used as input.

The results of two dataset given below are tabulated.

- Dataset RV11_BBS11022 from the mdsa_all version with four sequences.
- Dataset RV11_BBS11002 from the mdsa_100 version with eight sequences

Table 4: Comparison of Performance of Soga-Msa and Other Msa Tools

DATASET **RV11_BBS11022**				DATASET **RV11_BBS11002**			
MSA Tool	Exact Match	Alignment Length	Column Score (CS)	MSA Tool	Exact Match	Alignment Length	Column Score (CS)
Dialign	6	274	0.021	ClustalW	1	232	0.004
Mafft	11	208	0.052	Multalin	0	220	0
SOGA-MSA	**9**	**248**	**0.036**	**SOGA-MSA**	**1**	**259**	**0.003**

The results produced by SOGA-MSA and other tools like Dialign, Mafft, ClustalW and Multalin are tabulated above. It is observed that the CS of the alignment produced by SOGA-MSA is better than most commonly used tools like Dialign, Multalin and equal to ClustalW. The betterment of the multiple sequence alignment compliments the efficiency of the proposed algorithm.

VIII. Conclusion and Future Perspectives

In SOGA-MSA parameters like number of generations, chromosome length, crossover and mutation rate are made to adapt the values during execution, whereas in standard GA these values are determined before execution. In general, the values of various parameters of GA based algorithm are either default or selected from options. It is hard for a non-specialist to assign the values of various parameters without complete knowledge of the problem. Even default values may lead to bad results for some input. This is completely facilitated and proven by the proposed self-organizing approach of GA for MSA, where the parameter values are chosen by itself. The main advantage of SOGA-MSA is getting sequences alone as input from the user. Premature convergence considered as one of the major fitness range problems of standard GA is completely avoided by the execution for a range of rates.

The self-organizing crossover and mutation operator developed for MSA prevents the problem of repetition and increase of gaps in chromosomes. In addition, elitism selection avoids disruption of the best chromosome. The proposed SOBS self-organize the increase in mutation rate, which explores all rates within the range. As an advantage, this mechanism ensures that the best alignments produce for varying rates within the range are also included in the process of alignment.

Several widely used MSA tools like DCA[13] has a strong limitation in the number of sequences it can handle. In SOGA-MSA, there is no limitation in the number of input sequence and its length.

The algorithms used in other tools will produce the alignment with the same column score for every execution. However, in SOGA-MSA implementing the stochastic iterative algorithm, there is a chance of generating better alignments than the previous alignment in each execution. Further, it may generate better alignments with an increase in the number of generations also.

The future objectives are to self-organize other parameters like population size, crossover rate, etc., to minimize the execution time and to improve the quality of the alignment further.

References Références Referencias

1. T. D. Seeley, "When Is Self-Organization Used in Biological Systems?," *Biol. Bull.,* 2002, 202(3): 314–318.
2. J. H. Holland, "Adaptation in natural and artificial systems", University of Michigan Press, Ann Arbour, MI, 1975.
3. C. Gondro, B. P. Kinghorn, "A simple Genetic Algorithm for multiple sequence alignment", *Genet. Mol. Res.,* 2007, 6 (4): 964-982.
4. N. Kubota, T. Fukuda, K. Shimojima, "Virus-evolutionary genetic algorithm for a self-organizing manufacturing system", *Computers and Industrial Engineering,* 1996, 30(4): 1015-1026.
5. S. S. Ray, S. Bandyopadhyay, S. K. Pal, "New Genetic Operators for Solving TSP: Application to Microarray Gene Ordering", Springer-Verlag Berlin Heidelberg, 2005, pp. 605–610.
6. S. Diamantis,C. Anna, "Comparison of Multiple Sequence Alignment programs", M.Sc. Bioinformatics, National and Kapodistrian University of Athens.
7. C. Notredame, "Recent progresses in multiple sequence alignment: a survey", *Pharmacogenomics,* 2002, 3(1): 131-144.
8. J. D. Thompson, D. G. Higgins, T. J. Gibson, "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting position specific gap penalties and weight matrix choice", *Nucleic Acids Res.,* 1994, 22: 4673-4680. http://www.ebi.ac.uk/Tools/clustalw/
9. F. Corpet, "Multiple sequence alignment with hierarchical clustering", Nucleic Acids Res., 1988, 16: 10881-10890. http://bioinfo.genotoul.fr/multalin/multalin.html
10. B. Morgenstern, A. Dress, T. Wener, "Multiple DNA and protein sequence based on segment-to-segment comparison", *Proc. Natl. Acad. Sci.*, 1996, 93: 12098-12103. http://bibiserv.techfak.uni-bielefeld.de/dialign/
11. R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput", *Nucleic Acids Res.,* 2004, 32: 1792-1797. http://www.ebi.ac.uk/Tools/muscle/
12. C. Notredame, D. G. Higgins, J. Heringa, "T-Coffee: A novel method for fast and accurate multiple sequence alignment", *J Mol Biol.,* 2000, 302: 205-217. http://www.ebi.ac.uk/Tools/t-coffee/
13. J. Stoye, V. Moulton,A. W. Dress, "DCA: an efficient implementation of the divide-and conquer approach to simultaneous multiple sequence alignment", *Comput. Appl. Biosci.,* 1997, 13(6): 625-626. http://bibiserv.techfak.uni-bielefeld.de/dca/
14. C. Notredame, D. G. Higgins, "SAGA: sequence alignment by Genetic algorithm", *Nucleic Acids Res.,* 1996, 24(8):1515-1524.
15. K. Karadimitriou, D. H. Kraft, "Genetic Algorithms and the Multiple Sequence

Alignment Problem in Biology", In Proc. 2nd Annual Molecular Biology and Biotechnology Conference, Baton Rouge, LA, 1996.
16. S. A. Fatumo, I. O. Akinyemi, E. F. Adebiyi, "Aligning Multiple Sequence with Genetic algorithm", *International Journal of Computer Theory and Engineering,* 2009, 1(2): 179-182.
17. S. Wu, M. Lee, T. M. Gatton, "Multiple Sequence Alignment using GA and NN", International Journal of Signal Processing, *Image Processing and Pattern Recognition*, 21-30.
18. Introduction to genetic algorithm tutorial in www.obitko.com/tutorials/genetic-algorithms (Last accessed: 7.12.2010).
19. G. R. Harik, F. G. Lobo, "A Parameter-Less Genetic Algorithm", *IEEE Transactions on Evolutionary Computation*, 1999: 523-528.
20. T. Hong,H. Wang, W. Lin, W. Lee, "Evolution of Appropriate Crossover and Mutation Operators in a Genetic Process", *Applied Intelligence*, 2002, 16: 7–17.
21. H. Bao-Juan, Z. Jian, Y. De-Hong, " A Novel and Accelerated Genetic algorithm", *WSEAS Transactions on Systems and Control,* 2008, 3(4): 269-278.
22. R. Breukelaar, T. Bäck, "Self-Adaptive Mutation Rates in Genetic Algorithm for Inverse Design of Cellular Automata", July 2008: 12–16.
23. D. Thierens, "Adaptive mutation rate control schemes in genetic algorithms", Institute of Information and Computing Sciences, Utrecht Univerisity, The Netherlands, 2002.
24. J. Zhang, J. Zhuang, H. Du, S. Wang, "Self-organizing genetic algorithm based tuning of PID controllers", *Information Sciences*, 2009, 179 (7): 1007-1018.
25. J. T. Horng, E. M. Lin, B. H. Yang, E. Y. Kao, "A Genetic Algorithm for multiple sequence alignment", In Proc. of the GCB, 2001.
26. D. Liu, X. Xiong, Z. Hou, B. D. Gupta, "Identification of motifs with insertions and deletions in protein sequences using self-organizing neural networks", *Neural Networks,* 2005, 18 (5-6): 835-842.
27. H. Carroll, W. Beckstead, T. O'Connor, M. Ebbert, M. Clement, Q. Snell, D. McClellan, "DNA reference alignment benchmarks based on tertiary structure of encoded proteins", *Bioinformatics,* 2007, 23(19): 2648– 2649. http://dna.cs.byu.edu/mdsas/download.shtml

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

Approach to Job-Shop Scheduling Problem Using Rule Extraction Neural Network Model

By Mahmood Al Bashir, Md. Zahidul Islam, Dr. A. K. M. Masud
Bangladesh University

Abstract- This thesis focuses on the development of a rule-based scheduler, based on production rules derived from an artificial neural network performing job shop scheduling. This study constructs a hybrid intelligent model utilizing genetic algorithms for optimization and neural networks as learning tools. Genetic algorithms are used for obtaining optimal schedules and the neural network is trained on these schedules. Knowledge is extracted from the trained network. The performance of this extracted rule set is analyzed in scheduling a test set of 3x3 scheduling instances. The capability of the rule-based scheduler in providing near optimal solutions is also discussed in this thesis.

GJCST Classification: *I.2.6*

Strictly as per the compliance and regulations of:

Approach to Job-Shop Scheduling Problem Using Rule Extraction Neural Network Model

Mahmood Al Bashir, Md. Zahidul Islam, Dr. A. K. M. Masud

Abstract–This thesis focuses on the development of a rule-based scheduler, based on production rules derived from an artificial neural network performing job shop scheduling. This study constructs a hybrid intelligent model utilizing genetic algorithms for optimization and neural networks as learning tools. Genetic algorithms are used for obtaining optimal schedules and the neural network is trained on these schedules. Knowledge is extracted from the trained network. The performance of this extracted rule set is analyzed in scheduling a test set of 3x3 scheduling instances. The capability of the rule-based scheduler in providing near optimal solutions is also discussed in this thesis.

I. Introduction

Scheduling requires the arrangement of activities under constraints to meet a specific objective. A complex decision making activity it is, because of different conflicting goals, precise or limited resources and the difficulty in accurately modeling real world scenarios.

In today's highly competitive manufacturing environment, there is a distinct need for an integrated global approach towards production planning and control. The planning functions include demand forecasting, capacity and materials planning, process planning and operation scheduling. Scheduling theory is concerned with the mathematical formulation and study of various scheduling models and development of associated solution methodologies. The deterministic job shop scheduling problem (JSSP) consists of a finite set of jobs to be processed on a finite set of machines. The basic entity in the scheduling process is an operation, which refers to the processing of a particular job step on a specified machine. Various performance measures are used to evaluate the optimality of schedules ranging from minimization of makespan, tardiness and process cost to maximization of throughput and optimum resource utilization. JSSP is a constrained optimization problem (COP), where the precedence constraints on the problem are given by a predetermined order of operations for each job; and capacity or disjunctive constraints require that each operation be processed by only one machine at any given time. A schedule is the feasible resolution of the precedence and capacity constraints in the COP [1].

About- Department of Industrial and Production Engineering (IPE), Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh

The current research uses Artificial Neural Networks (ANNs) as the machine learning tool of choice to study the scheduling process. ANNs are being recognized as a powerful and general technique for machine learning because of their non-linear modeling abilities. Further, their distributed architecture is more robust in handling the noise-ridden data. The hypothesis or model learned by the neural network is not explicitly stated, but is implicitly enumerated in the network architecture. However, ANNs can be made to yield comprehensible models by using rule extraction procedures. This thesis has three major objectives:

- To train an ANN on the schedules generated by a GA, to predict the priority of an operation in a schedule based on the job attributes.
- To capture the embedded knowledge by extracting symbolic rules and decision trees by using appropriate ANN rule extraction algorithms.
- A comparative evaluation of the predictive accuracy of the extracted rule set, the trained ANN, GA scheduler and other machine learning algorithms.

II. Approach

The deterministic job shop scheduling problem (JSSP) is one of the classical problems in scheduling literature. JSSP consists of a finite set of *n* jobs to be processed on a finite set of *m* machines and is denoted as an *n* x *m* problem. The routing of a job is a predetermined sequence of operations. Each operation is processed on a specified machine and has a predetermined processing time. The job routings and the associated processing times are given by a definite process plan. JSSP is a constrained optimization problem (COP) where the precedence constraints on the problem are given by the job routings. Capacity or disjunctive constraints require that each operation be processed by only one machine at any given time. Other assumptions include the following:

- Machine repetitions by a job are not allowed.
- Machine absences are not allowed (i.e., each job is processed on every machine).
- Uninterrupted processing of operations without preemption.
- No machine breakdowns throughout the scheduling process.

- Transportation time between machines is zero.
- The job shop is static and deterministic in nature i.e., there is no randomness involved in determining all the necessary parameters for defining the job shop problem.

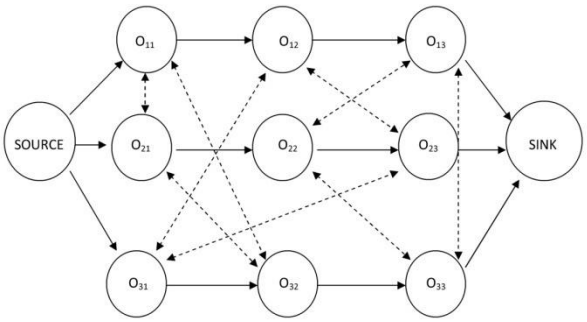


Figure1: The disjunctive graph representation of the 3 x 3 problem. Adapted from Yamada and Nakano [2]

In the above figure, the conjunctive constraints are given by complete arrows and the dashed arrows indicate the disjunctive constraints. Two fictitious nodes, source and sink nodes are added to the graph to represent the starting and ending operations. Panwalker and Iskander [3] provide a comprehensive survey of scheduling heuristics. The main drawback in using these elementary rules is that different rules perform best in different scenarios and no single rule dominates the rest across all scenarios. To improve performance, probabilistic combinations of the elementary priority rules are often employed for determining priority. Blackstone *et al.* [4] provide a detailed comparison of several elementary dispatching rules and their combinations. Lawrence [5] compares the performance of ten individual priority dispatch rules with a randomized combination of these rules. Superior results were delivered by the combination method, but it required substantially more computing time. Kaschel *et al.* [6] provide empirical results of the performance of priority rules in scheduling several benchmark instances. The authors compare single priority rules, simple combinations of them and combinations of priority rules by the Analytic Hierarchy Process method. AHP is a statistical decision making tool capable of generating weighted combinations of priority rules and provided the best results in this study.Most symbolic Artificial Intelligence (AI) approaches cast the JSSP as a constraint satisfaction problem. A CSP specifies a set of decisions to be made and a set of constraints to determine the validity of such decisions. The general procedure to solve a CSP is to reduce the search space by utilizing a constructive search strategy. Such a strategy incrementally builds a solution by assigning values to variables and checking for constraint violations. If any violations are found, a backtracking strategy is employed to undo previous variable assignments. The procedure is repeated with a fresh set of variable assignments. The Intelligent Scheduling and Information System constructed by Fox [7] is a good example of the AI scheduling system. There are many variations of the generic constraint satisfaction procedure. Fox and Sadeh [8] provide a comparative summary of a variety of constraint satisfaction approaches applied to a set of benchmark scheduling instances.

Neural network scheduling systems offer an alternate AI-based scheduling paradigm. Cheung [9] provides a comprehensive survey of the main neural network architectures used in scheduling. These are: searching network (Hopfield net), probabilistic network (Boltzmann machine), error-correcting network (multilayer perceptron), competing network and self-organizing network. Jain and Meeran [10] also provide an investigation and review of the application of neural networks in JSSP.

III. Methodology

The knowledge base for the learning task was provided by the genetic algorithm's solution to the job shop problem. For this purpose, a well-known 3x3 problem instance, ft03 devised by has been chosen as the benchmark problem. This test instance has three jobs, each with three operations to be scheduled on three machines and has a known optimum makespan of 11 units. The data for the instance is shown in Table 1 using the following structure: machine, processing time.

Table 1 The ft03 instance

Job	Operation		
	1	2	3
1	1,1	2,4	3,8
2	2,7	3,4	1,11
3	1,12	2,5	3,2

The schedules obtained by the GA contain valuable information relevant to the scheduling process. The learning task was to predict the position of an operation in the sequence, based on its features or attributes. Based on a study of operation attributes commonly used in priority dispatch rules, the following attributes have been identified as input features: operation, process time, remaining time and machine load. These input features have been clustered into different classes using the concept hierarchy for 6 x 6 job shop problems developed by Koonce and Tsai [11].
Operation: Each job has three operations that must be processed in a given sequence. The *Operation* feature identifies the sequence number of the operation ranging between 1 and 3. This feature has been clustered into three classes as: *First*, *Middle* and *Last*.
ProcessTime and RemainingTime: The *ProcessTime*feature represents the processing time for

the operation. The *RemainingTime*feature denotes the sum of processing times for the remaining operations of that job and provides a measure of the work remaining to be done for completion of the job. For the benchmark ft03 instance, the processing times ranged from 1 to 4 units, while the remaining times ranged from 0 to 18 units. Based on the data, three classes (clusters) for these features were identified as follows. The ranges were split into three equal intervals. The first interval was classified as Short, second as Medium, and last interval was labeled as *Long*. Table 2 shows the classification of the *ProcessTime*and *RemainingTime*for these intervals.

Table 2 ProcessTime,RemainingTime feature classes

Attributes	Short	Middle	Long
ProccessTime	[1, 4]	[5, 8]	[9, 12]
RemainingTime	[0, 6]	[7, 12]	[13,18]

Machine Load: The *Machine Load* feature determines the machine loading and was clustered into two classes: *Light and Heavy*. This feature represents the capacity or utilization of machines in units of time and Table3 shows the classification of machine loading for the ft03 instance.

Table 3 MachineLoad feature classification

Attributes	Light	Heavy
Time	Less than 7	Greater than 7

The machine processing times range between 1 and 12 units, with an average of 6.5 units. All the machines having processing times less than 7 units were classified as *Light*, while those having processing times greater than 7 units were labeled as *Heavy*.

Priority: The target concept to be learned was the priority or position in the sequence. Since an operation can be positioned in any one of the 9 locations available in the sequence, it may be difficult to discover an exact relationship between the input features and the position. However, if the problem was modified to predict a range of locations for the operation, the learning task becomes easier. The target feature priority, thus determines the range of positions in the sequence where the operation can be inserted. The possible range of positions have been split into 4 classes and assigned class labels as shown in Table 4.

Table 4 Assignment of class labels to target feature

Range of Position	Priority
0 – 1	One
1 – 3	Two
4 – 6	Three
7 – 9	Four

There are three aspects related to development of a neural network model. The first is the choice of the training, cross-validation (CV) and testing data sets and their sizes, the second is the selection of suitable architecture, training algorithm and learning constants, and the third is the determination of the termination criteria. Unfortunately, there are no definitive heuristics or formulae to determine these parameters. Considerable experimentation was necessary to achieve a good network model of the data. The software NeuroShell Predictor developed by Ward System Group, Incorporated was used for development and testing of the neural network model.

Training, Cross-validation and Test Datasets: The 144 optimal schedules obtained by the GA represent a total number of 108 operations (12 schedules x 9 operations/schedule). Assignment of input features and target classes was done for each operation according to the classification scheme described in the previous subsection. Sample data for the classification task is shown in Table5.

Table 5 Sample data for the classification task

Pattern_ID	Operation	Process Time	Remaining Time	Machine Load	Priority
1	First	Short	Short	Low	1
2	Mid	Mid	Mid	Low	2
. . .	. . .	. . .	. . .	. . .	. . .
1,295	Last	Mid	Mid	High	3
1,296	Last	Long	Long	High	4

The entire data set included 24 distinct input patterns with different target feature values (priority), constituting a total of 1,296 patterns (exemplars). This classification data set was split into training, cross validation and testing data sets with 70%, 15% and 15% memberships.

The neural network can be considered an implicit model of the training data. The goal of the rule extraction algorithms is to translate this implicit model into explicit symbolic form. In this work, the extracted knowledge is captured in two symbolic representations: decision trees and propositional rules. The possible rule space for these procedures is derived in the following way. The number of classes in the input features of the classification problem(Operation, ProcessingTime, RemainingTime and the MachineLoad features) are three, three, three and two respectively. Hence, the number of possible rule antecedent combinations (patterns) in the rule space is 3 x 3 x 3 x 2 = 54.

In this rule set, referred to as NN-Rule set, the keyword “Any” denotes all possible values of an attribute. The first rule in NN-Rule set implies that the *First* operation for a job with a *Short* processing time, a *Short* remaining time and processed on a machine of

any load class (i.e., *Light* or *Heavy*) has an associated priority of *one* (highest priority) for scheduling in the sequence.

A candidate rule is verified based on the convergence of the forward and backward passes. A run constitutes one forward and one backward pass. The procedure terminates when the difference between validity intervals between two consecutive runs becomes less than a predefined tolerance (a small threshold). Generally, the number of runs required for convergence can be considered a function of the specified validity intervals and the parameters of the neural network (weights and biases). A contradiction in the above procedure implies that the candidate rule incorrectly describes the behavior of the neural network. Such a rule is expunged from the candidate rule set and the procedure is repeated with the other candidate rules. This process continues until the candidate rule set is exhausted. A tolerance of 0.01 was chosen for termination. All the 36 rules in the NN-Rule set were verified by this procedure. Among the 1296 data, 1000 data were used to train the network and the rest were used to test the network. A total of 80 hidden neurons were trained with a optimal of 16 hidden neurons. The network performance was at an optimum rate of 0.8102. The lack of the training can be attributed to the amount of data used in the training procedure. Had more data were used, the best network performance could have been better.

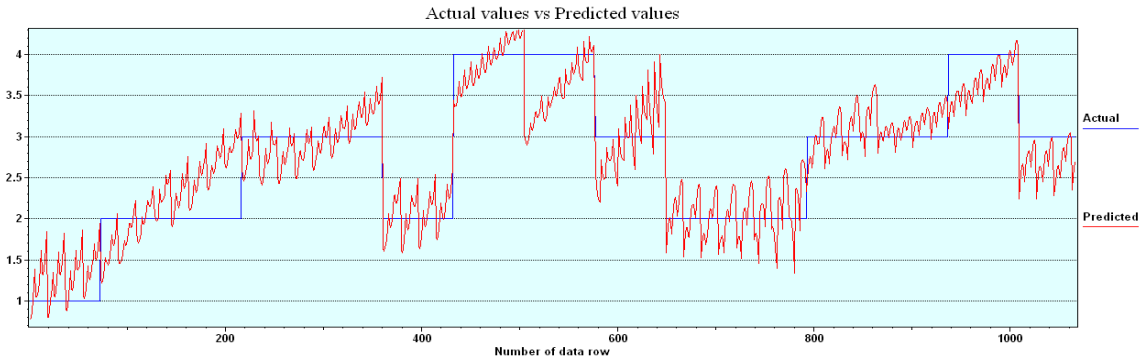


Figure 2: Actual vs. Predicted value graph in training phase (Enlarged in appendix C)

The efficacy of the rule extraction task was evaluated along the following dimensions:

- Comprehensibility and Expressive Power: The propositional rules in the NN-Rule set developed by the rule extraction procedures used easily understood feature and class labels for describing the extracted knowledge. The antecedent of each rule in the rule set was a simple conjunction of the input features. The number of input features in the antecedent of a rule provides an indirect measure of the comprehensibility and expressive power of the rule set. Table 5.3 shows the rules in the NN-Rule set (shown on the right) segregated according to the number of features in their respective antecedents (shown on the left). Approximately, a third of the rules had fewer than four antecedents (maximum) increasing the comprehensibility of the developed rule set.

Table 6 Number of features in the rule antecedent for the NN-Rule set

Number of features in the rule antecedent	Number of rules (Total 36)
1	None
2	2
3	11
4	23

- Accuracy and Fidelity: The rule set accurately mimicked the behavior of the trained neural network in classifying all the patterns in the rule space. Hence, the fidelity of the extraction process was maximum.

To schedule the 3 x 3 benchmark instance (ft03), a priority index was assigned to each of the 36 operations. This was accomplished by matching the features of the operation with the antecedent of the induced rules in the NN-Rule set. The consequent of the matched rule represented the priority index of that operation. This algorithm chooses from among the available operations based on the priority index. Also, the operations were locally left-shifted to improve the makespan of the generated schedule. The Gantt chart was used as a tool for visualizing the developed schedules. A similar procedure was utilized for scheduling the problem with the ID3-Rule set and the Shortest Processing Time (SPT) heuristic.

After the completion of the training phase the rest of the data were used to test the neural network to find the optimum level of usage it can offer. With the rest 296 data the network was run to test the difference between the actual and predicted value. An average error of .03182 was originated which yielded the network to be efficient enough to be used comprehensibly. The

data obtained from a manufacturing facility in Bangladesh is also decided to be tested in this network with a view to finding the average error the network shows and the stability of the network.

IV. Conclusion

This paper presents a novel knowledge-based approach for the job shop scheduling problem by utilizing the various constituents of the soft computing paradigm. The ability of a genetic algorithm (GA) to provide multiple optimal solutions was exploited to generate a knowledge base of good solutions. A neural network was successfully trained on this knowledge base. Then, rule extraction algorithms were employed to induce decision tree and propositional rule representations describing the behavior of the trained neural network. The rule extraction task was successful in generating a rule set which completely and accurately mimicked the behavior of the trained neural network. The scheduler developed from this rule set can be utilized to schedule any 3 x 3 job shop scenario. Also, the developed system provides knowledge in the form of comprehensible rules which can effectively aid a human in the scheduling task.

A test problem set consisting of 10 randomly generated 3 x 3 scenarios was used to evaluate the performance of the developed rule-based scheduler. The makespans produced by the GA were considered to be the known optimal solutions for these scenarios. The rule-based scheduler had a deviation of 4.6 time units (8.4 %) from the optimum (i.e., average makespan of the GA) on the test problem set. Also, the rule-based scheduler performed better than the Shortest Processing Time (SPT) heuristic in all ten cases. Though the rule-based scheduler could not match the performance of the genetic algorithm, it is computationally less intensive than the GA and offers a more comprehensible scheduling approach. It also provides an attractive alternative to simple heuristics like SPT for scheduling 6 x 6 job shop problems.

A comparative evaluation of the rule-based scheduler with other schedulers developed from different machine learning methodologies was also undertaken. Two schedulers developed by other researchers using the Attribute-Oriented Induction (AOI) data mining methodology and another scheduler based on the ID3 decision tree induction algorithm were used for comparison. Among these schedulers, the rule-based scheduler developed in the current work had the closest average makespan to that of the genetic algorithm. However, statistical analysis revealed no significant differences in the performance of these schedulers on the test problem set.

The similar performance of the current approach compared to the AOI data mining methodology proves the feasibility of neural network based data mining. Unlike the rule set derived in this work, those induced from AOI and ID3 methods were insufficient to describe any randomly generated 3 x 3 scenario. Also, the decision tree induction algorithm utilized for knowledge extraction from the neural network is similar to the ID3 algorithm. The deviation in their performance of the rule-based scheduler and the ID3-based scheduler is mainly attributable to the robustness of the neural networks in handling noisy data sets.

In summary, this research was able to successfully develop a rule-based scheduler, which provides a close approximation to the performance of a GA scheduler for the 3 x 3 job shop scheduling problems.

This research focused primarily on deriving production rules from a neural network performing job shop scheduling. There is a definite scope for improvement in the current research along the following directions.

Use of multiple data sets

The knowledge base for training the neural network in the current approach was derived from solutions to a single benchmark 3 x 3 problem. This knowledge base can be augmented with near-optimal solutions to randomly generated 3 x 3 scenarios provided by a GA. This can lead to an improvement in the generalization capabilities of the trained neural network.

V. Acknowledgemetnt

The authors are heartily thankful to their supervisor, Professor Dr. A. K. M. Masud, Professor and Head, Department of Industrial and Production Engineering, Bangladesh University of Engineering and Technology, whose spirit to work, guidance and support from the initial to the final level enabled them to develop an understanding of the subject. Above all and the most needed, he provided them unflinching encouragement and support in every possible ways.

The authors express their profound thanks and gratefulness to ShuvaGhosh, Assistant Professor, Department of Industrial and Production Engineering, Bangladesh University of Engineering and Technology for his advice, supervision, and crucial contribution, which made the thesis possible to initiate.

Finally, the authors would like to thank everybody who were important and helped them out in every process to the successful realization of thesis.

References Références Referencias

1. Baker, K. (1974). Introduction to sequencing and scheduling. New York, NY: John Wiley & Sons, Inc.
2. Yamada, T., & Nakano, R. (1995). Job shop scheduling by simulated annealing combined with deterministic local search. Metaheuristics

International Conference. Hilton, Breckenridge, Colorado, USA, 344-349.
3. Panwalker, S. S., &Iskander, W. (1977). A survey of scheduling rules. Operations Research, 25, 45-61.
4. Blackstone Jr., J. H., Phillips, D. T., & Hogg, G. L. (1982). A state-of-the-art survey of dispatching rules for manufacturing job-shop operations. International Journal of Production Research, 20, 27-45.
5. Lawrence, S. (1984). Supplement to resource constrained project scheduling: An experimental investigation of heuristic scheduling techniques. Graduate School of Industrial Administration.Carnegie-Mellon University, Pittsburgh, USA.
6. Käschel, J., Teich, T., Köbernik, G., & Meier, B. (1999). Algorithms for the job shop scheduling problem: A comparison of different methods. European Symposium on Intelligent Techniques. Greece, June 3-4.
7. Fox, M. S. (1987). Constraint-directed search: A case study of job-shop scheduling. Research Notes in Artificial Intelligence. London: Pitman Publishing.
8. Fox, M. S., &Sadeh, N. (1990). Why is scheduling difficult ?A CSP perspective. In Aiello, L. (ed), ECAI-90 Proceedings of the 9th European Conference on Artificial Intelligence. Stockholm, Sweden. August 6-10, 754-767.
9. Cheung, J. Y. (1994). Scheduling. In Dagli, C. H. (ed), Artificial Neural Networks for Intelligent Manufacturing. London: Chapman and Hall, Chapter 8, 159-193.
10. Jain, A. S., &Meeran, S. (1998). Job-shop scheduling using neural networks. International Journal of Production Research, 36(5), 1249-1272.
11. Koonce, D. A., & Tsai, S. C. (2000). Using data mining to find patterns in genetic algorithm solutions to a job shop schedule. Computers & Industrial Engineering, 38 (3), 361-374.

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

Future Biometric Passports and Neural Networks

By Kheder Durah, Omar Al Zoubi, Bilal Alzoubi, Emad Mohammed

Abstract- Due to the increase in the number of crimes and different ways they are perpetrated, demand has increased on the means that increase the level of security accuracy in the places that need special kind of protection, and places that require verifying the identity of those who demand access, such as computer networks, banks and home land security departments. There are many ways to identify people and grant them the required access; these methods include: What people have? (like an access card or key) and What people know? (like password); Moreover, there are physical biometric features such as (figure prints, retina, iris, DNA,etc) and behavioral biometric features such as (signature, voice, walking, etc). Recently, experience proved that using the iris is the best and more accurate than any other way and it will be the target of our research. There are several ways to increase the level of security that have been innovated, most important of which was using the biometrics. The most accurate biometric feature is the human eye iris, due to the characteristics it enjoys, and which make it possible to be used to identify people. The eye iris texture differs from one person to another; it even differs between identical twins, and the right and left eyes of the same person too. The aim of this research is to design an algorithm to recognize the iris for using it to identify people and create an international biometric passport for that person.

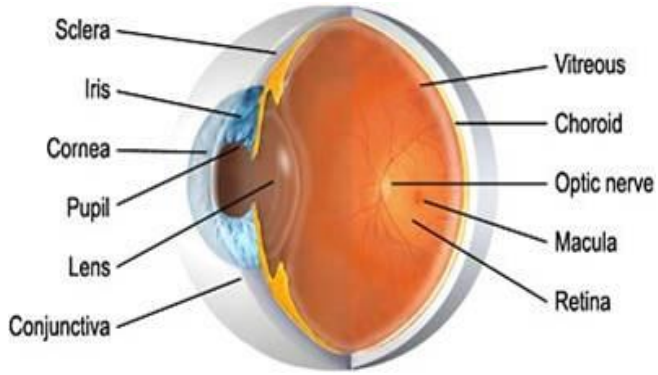
GJCST Classification: I.2.6

Strictly as per the compliance and regulations of:

Future Biometric Passports and Neural Networks

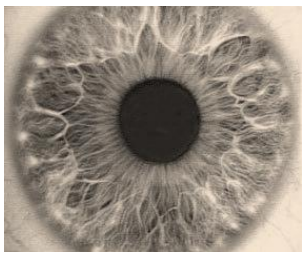
Kheder Durah, Omar Al Zoubi, Bilal Alzoubi, Emad Mohammed

Abstract– Due to the increase in the number of crimes and different ways they are perpetrated, demand has increased on the means that increase the level of security accuracy in the places that need special kind of protection, and places that require verifying the identity of those who demand access, such as computer networks, banks and home land security departments. There are many ways to identify people and grant them the required access; these methods include:
What people have? (like an access card or key) and What people know? (like password); Moreover, there are physical biometric features such as (figure prints, retina, iris, DNA,etc) and behavioral biometric features such as (signature, voice, walking, etc). Recently, experience proved that using the iris is the best and more accurate than any other way and it will be the target of our research. There are several ways to increase the level of security that have been innovated, most important of which was using the biometrics. The most accurate biometric feature is the human eye iris, due to the characteristics it enjoys, and which make it possible to be used to identify people. The eye iris texture differs from one person to another; it even differs between identical twins, and the right and left eyes of the same person too. The aim of this research is to design an algorithm to recognize the iris for using it to identify people and create an international biometric passport for that person.



I. Introduction

Several methodologies using the iris for identifying people have been used, most important was that applied by scientist J. Daugman. Generally speaking, all those are similar in their steps of operation; what is different in our research is the way of performing each, some were modified in one step while others were done in more than one step.



The process of identifying and recognizing people using the eye iris goes through the following steps:

1. Iris image acquisition
2. Processing
3. Post processing
4. Matching
5. Decision

We carried out few thousands lab tests and experiments and took several pictures of the eye iris by a special IRIS capture camera used for this purpose namely Panasonic Authenticam using Iridian technologies; We also developed a computer program using MATLAB language to code the iris and execute the matching process between one iris and another. In this research paper we developed new ways in many stages:

- We used a new method to locate inner and outer boundaries of the iris by considering that the iris is not a circle, it has an elliptical shape. We used a method to locate the iris boundaries by generating circles with different diameters and different centers, and then we measure the mean of the intensity for each circle.
- We used a new method t select a sector of the iris by locating the cross points of the eyelashes with the iris.
- We used neural networks based on simplex algorithm to verify the person who stand in front of the camera.

II. Why Choosing Eye Iris

The most accurate biometric feature is the human eye iris, due to the characteristics it enjoys, and which make it possible to be used to identify people. The eye iris texture differs from one person to another; it

even differs between identical twins, and the right and left eyes of the same person too.[1], [4].

Iris characteristics can be summarized in the following:

- Iris is more unique than other biometric features like (fingerprints, retina, hand, face, voice, walk, etc).
- Iris capture is non-intrusive and does not cause any inconvenience to people unlike the DNA which is highly accurate but intrusive and may cause discomfort and violation of privacy.
- Iris is an internal organ and well protected from external dangers.
- Iris is fixed since early infantry and throughout person's life.
- Iris capture procedure can be performed from a suitable convenient distance.
- Iris is not affected by genetics.
- Iris signature can be easily obtained, encoded and used in digital environment.

There are also some challenges for using the Iris for identifying people:

- The object to deal with (iris) is small.
- The dynamic movement of Iris which requires extra technical efforts to capture the right angle of iris.
- Iris is positioned behind a curved and reflective surface namely cornea.
- We may not get a good capture of the iris due to eyelashes, eyebrows and contact lens.
- Iris size changes with light source. It gets wider in dark and smaller in light.
- To have a perfect iris image, the light source must be hidden and not shiny.

III. Application of the Iris in Biometric Systems

Iris recognition has tremendous potential for security in any field. The iris is extremely unique and cannot be artificially impersonated by a photograph (Daugman, 2003).This enables security to be able to restrict access to specific individuals.

An iris is an internal organ making it immune to environmental effects. Since an iris does not change over the course of a lifetime, once an iris is encoded it does not need to be updated. The only drawback to iris recognition as a security installment is its price, which will only decrease as it becomes more widely used.

A recent application of iris recognition has been in the transportation industry, most notably airline travel. The security advantages given by iris recognition software have a strong potential to fix problems in transportation (Breault, 2005). Its most widely publicized use is in airport security. IBM and the Schiphol Group engaged in a joint venture to create a product that uses iris recognition to allow passengers to bypass airport security (IBM 2002). This product is already being used in Amsterdam. A similar product has been installed in London's Heathrow, New York's JFK, and Washington's Dulles airports (Airport, 2002). These machines expedite the process of passengers going through airport security, allowing the airports to run more efficiently.

Iris recognition is also used for immigration clearance, airline crew security clearance, airport employee access to restricted areas, and as means of screening arriving passengers for a list of expelled persons from a nation (Daugman, 2005). This technology is in place in the United States, Great Britain, Germany, Canada, Japan, Italy, and the United Arab Emirates.

IV. Biometric System Stages

The system has the following stages [8]

a) *Image acquisition*

We use a special camera to capture a clear picture of the eye, in order to have a clear shot of the inner eye boundaries with the pupil and the outer boundaries with eye Sclera.

b) *Preprocessing*

In this stage several activities takes place to process the image. We shall try to summarize them as follows:

a) Convert the color RGB image to Gray scale.
b) Normalize of the image's histogram in order to distribute the density on the entire range [0-256].
c) Images dimensions are unified.
d) Determine the inner and outer boundaries of the eye iris. In this step, the determination of the inner boundaries of the Iris is identified; with the assumption that the pupil position from the outer boundaries compared to the inner boundaries of the Iris with a rate of 1.5 from the eye center as shown in the following figure:

We may also add a constant value related to the race of the human (white, black, and yellow) as we have noticed from the many thousands of experiments conducted to this respect.

e) Iris center is then determined assuming that it is equal to the center of the pupil, although in many cases we found that both centers may have a difference between 5-6 pixels, but such small difference will affect our results.
f) Then we convert Image's Cartesian coordinates to Polar coordinates using the following method:

- With the assumption that the center of pupil is identical of the Iris center, we draw x, y access at the center of the image.
- We assumed that the cross section of these lines with the outer boundaries forms the frame edges where the picture is placed within.
- Then we convert the Cartesian coordinates to Polar coordinates based on the center of the image and not based on the center of the pupil.
- Then we remove the pupil area because it contains no information, and hence we are left with a clear iris.

In general, we can convert Cartesian coordinates to Polar coordinates using the following equations:

$$I\left(x(p,\theta), y(p,\theta)\right) = I(p,\theta)$$
$$x(p,\theta) = (1-p)\,.\,Xp(\theta) + p\,.Xi(\theta)$$
$$y(p,\theta) = (1-p)\,.\,Yp(\theta) + p\,.Yi(\theta)$$

Whereas

$$Xp(\theta) = Xp\theta(\theta) + Rp\,.\cos(\theta)$$
$$Yp(\theta) = Yp\theta(\theta) + Rp\,.\sin(\theta)$$
$$Xi(\theta) = Xi\theta(\theta) + Ri\,.\cos(\theta)$$
$$Yi(\theta) = Yi\theta(\theta) + Ri\,.\sin(\theta)$$

With
[(2π,0)θ]θ and p[p(1,0)] represents the parameters that describe the conversion system to Polar coordinates.
Rp , *Ri* represents the radius of the pupil and iris respectively; and
Yi(θ) , Xi(θ) and p(θ),Xp(θ) represents the iris and pupil boundaries coordinates respectively.

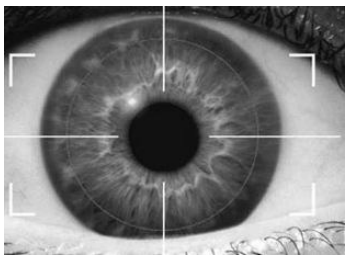


Figure (1) Original Image - the vertical and horizontal lines crossing the eye pupil at the capture stage

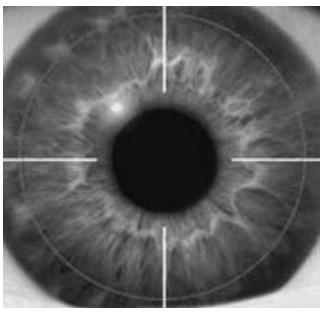
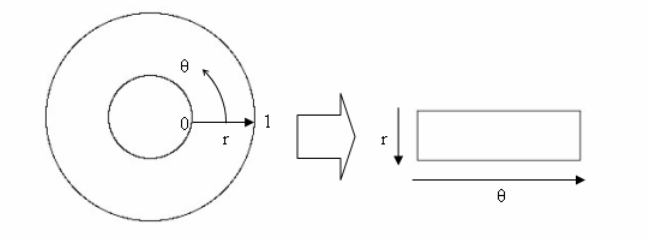


Figure (2) iris image after determining its inner and outer boundaries (outlines located)



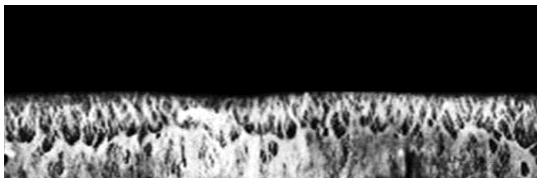


Figure (3) the iris image after conversion from Cartesian to Polar coordinates (Re-sampled Polar-Cartesian)

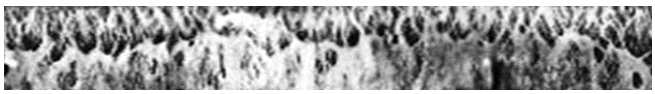


Figure (4) the iris image that will be processed and encoded (Intensity Enhanced)

c) *Post-processing stage.*

In this stage we encode the iris in order to save it at the end of the entry as a template. This process can be performed by the application of Gabor filter [8] or Wavelet Transformation [2], [3].

Application of Gabor Filter:

- We apply Gabor filer on the iris part and we get at the end we get two values that determines the image's elements. One of these values is real and the other is imaginary.
- Then we extract the phase for each value and prepare them for encoding.
- At this stage we encode the phases assuming "1" for values greater than 1 and "0" for values less than in order to get final values representing the phase.

Application of Wavelet Transformation

The aim of using this method is to have few images for the eye in different resolutions. Wavelet is cable of dealing with the smallest details in the image which is an advantage compared with other transformation methods.

* When application of Wavelet, the image passes through two types of filters, high and low frequency as shown in Figure (5):

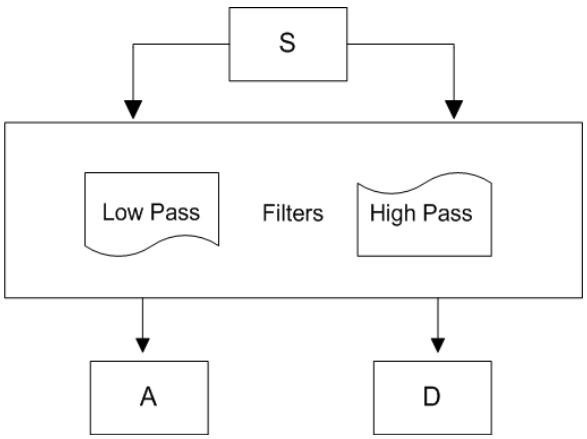


Figure (5): First stage in application of Wavelet transformation

We get two coefficients namely: "A" representing low frequencies in the signal called approximation coefficients, and we get "D" representing high frequencies in the signal called detailed coefficients; This coefficients also has detailed confidents namely (H) horizontal, (V) vertical and (D) diagonal.

* We then pass these approximate coefficients in to high and low frequencies filters in order to get new set of coefficients namely A1 and D1; and so on we repeat the process several times. In ideal situations when dealing with images the process is repeated /3/ times and when dealing with uni-directional frequencies, the process is repeated /7/ times.

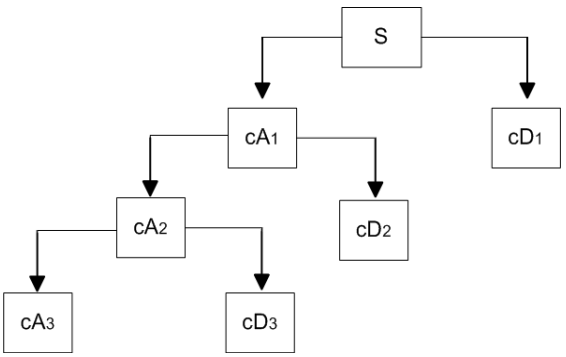


Figure (6): 3-level Wavelet transformation

* We can also apply the same analysis on the detailed coefficients until we get a magnified signal analysis.

* The next important step is to choose the best filter as there are many of them:
  - Haar
  - Daubechies
  - Biorthogonal
  - Coiflets
  - Symlets
  - Morlet
  - Mexican Hat
  - Meyer

In this research paper, we shall use "Haar" for simplicity; and as we apply this filter we get various different images of the same iris in different resolutions; we then combine the D and V coefficients for the last level of the analysis in one matrix. After that we add the H coefficient to the matrix to get a final one which will be encoded as a template for the original iris used for future comparisons.

d) *Matching Stage*

This process is carried out using the stored template and the captured iris using the hamming distance given in the following formula:

$$HD = \frac{1}{N}\sum_{j=1}^{N} Xj(XOR)Yj$$

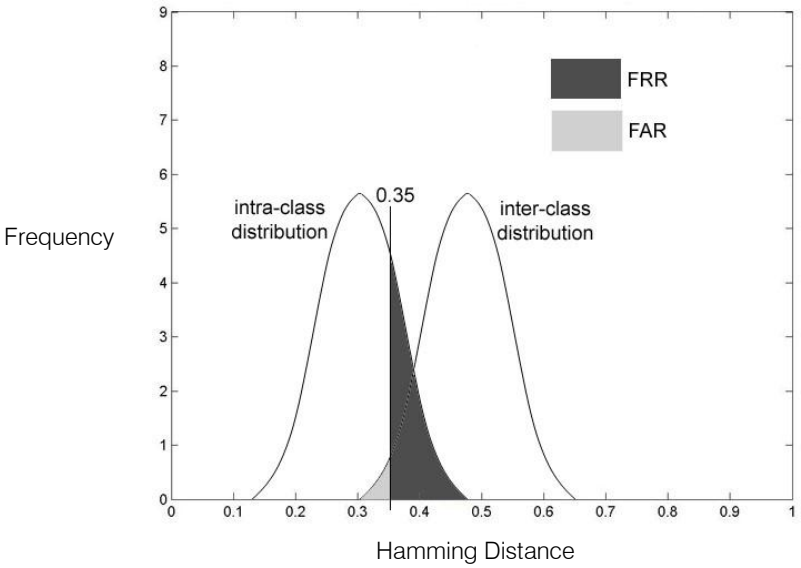
Whereas,
"N" is the number of elements in the matrix.
"X,Y" are the elements in the both matrixes being compared.

The matching is considered to be identical if HD = 0 and if HD = 1 then there is no match. By matching we mean deciding the identity of the person subject to the test.

e) *The Decision Stage*

In this stage we determine the Threshold of the matching value which will highly affect the nature of the decision the system will carry out. This stage is also critical as it also determine the fault rates of the system namely the False Accept (FAR) and False Reject Rates (FRR).

In the system that was designed by "J. Daugman, 2003" the Threshold was 0.32, meaning if HD < 0.32 then we have a match, and if HD > 0.32 then we don't have a match.



References Références Referencias

1. Kheder Durah, ICT Consultant, Associate Professor and Deputy Dean at Faculty of Computer Science and Engineering, Taibah University, Yanbu Branch, Kingdom of Saudi Arabia, email: kdurah@gmail.com, Tel: (+966)552658076.
2. Omar Al Zoubi, Assistant Prof. and Dept. Head at Faculty of Computer Science and Engineering, Taibah University, Yanbu Branch, Kingdom of Saudi Arabia, email: dr_omar1978@yahoo.com, Tel: (+966)530634650.
3. Bilal Alzoubi, Assistant Prof. Faculty of Computer Science, OmAlqora University, ALQunfdah, Kingdom of Saudi Arabia, email: bilal_14879jo@yahoo.com , Tel: (+966)507233770.
4. S. Sanderson, J. Erbetta. Authentication for secure environments based on iris scanning technology. *IEE Colloquium on Visual Biometrics*, 2000.
5. J. Daugman. How iris recognition works. Proceedings of 2002 International Conference on Image Processing, Vol. 1, 2002.
6. E. Wolff. *Anatomy of the Eye and Orbit*. 7th edition. H. K. Lewis & Co. LTD, 1976.
7. R. Wildes. Iris recognition: an emerging biometric technology. *Proceedings of the IEEE*, Vol. 85, No. 9, 1997.
8. J. Daugman. Biometric personal identification system based on iris analysis. United States Patent, Patent Number: 5,291,560, 1994.
9. J. Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 11, 1993.
10. R. Wildes, J. Asmuth, G. Green, S. Hsu, R. Kolczynski, J. Matey, S. McBride. A system for automated iris recognition. *Proceedings IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, pp. 121-128, 1994.
11. W. Boles, B. Boashash. A human identification technique using images of the iris and wavelet transform. *IEEE Transactions on Signal Processing*, Vol. 46, No. 4, 1998.
12. S. Lim, K. Lee, O. Byeon, T. Kim. Efficient iris recognition through improvement of feature vector and classifier. *ETRI Journal*, Vol. 23, No. 2, Korea, 2001.
13. S. Noh, K. Pae, C. Lee, J. Kim. Multiresolution independent component analysis for iris identification. *The 2002 International Technical Conference on Circuits/Systems, Computers and Communications*, Phuket, Thailand, 2002.
14. Y. Zhu, T. Tan, Y. Wang. Biometric personal identification based on iris patterns. *Proceedings of the 15th International Conference on Pattern Recognition*, Spain, Vol. 2, 2000.
15. C. Tisse, L. Martin, L. Torres, M. Robert. Person identification technique using human iris recognition. *International Conference on Vision Interface*, Canada, 2002.
16. Chinese Academy of Sciences – Institute of Automation. *Database of 756 Greyscale Eye Images*. http://www.sinobiometrics.com Version 1.0, 2003.
17. C. Barry, N. Ritter. *Database of 120 Greyscale Eye Images*. Lions Eye Institute, Perth Western Australia.

18. W. Kong, D. Zhang. Accurate iris segmentation based on novel reflection and eyelash detection model. *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, 2001.
19. L. Ma, Y. Wang, T. Tan. Iris recognition using circular symmetric filters. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 2002.
20. N. Ritter. Location of the pupil-iris border in slit-lamp images of the cornea. *Proceedings of the International Conference on Image Analysis and Processing*, 1999.
21. M. Kass, A. Witkin, D. Terzopoulos. Snakes: Active Contour Models. *International Journal of Computer Vision*, 1987.
22. N. Tun. *Recognising Iris Patterns for Person (or Individual) Identification*. Honours thesis. The University of Western Australia. 2002.
23. D. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, 1987.
24. P. Burt, E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*. Vol. 31 No. 4. 1983.
25. P. Kovesi. *MATLAB Functions for Computer Vision and Image Analysis*.Available at: http://www.cs.uwa.edu.au/~pk/Research/MatlabFns/index.html
26. A. Oppenheim, J. Lim. The importance of phase in signals. *Proceedings of the IEEE 69*, 529-541, 1981.
27. P. Burt, E. Adelson. The laplacian pyramid as a compact image code. *IEE Transactions on Communications*, Vol. COM-31, No. 4, 1983.
28. J. Daugman. Biometric decision landscapes. *Technical Report No. TR482, University of Cambridge Computer Laboratory,* 2000.
29. T. Lee. Image representation using 2D gabor wavelets. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, Vol. 18, No. 10, 1996.
30. Birkhauser, "Friendly Guide to Wavelet", Boston 1994.
31. MATLAB 5, source code for the iris recognition software presented in this publication is available via the World Wide Web at the address. http://www.csse.uwa.edu.au/~masekl01/.
32. J. Daugman, "Anatomy and Physiology of the Iris", 2003 from www.cl.cam.ac.uk/users/jgd/addisadvans.html.
33. European Commission Joint Research Center (DG JRC) Institute for progressive technologies studies, 2005 www.jrc.es
34. J. Daugman, "How Iris Recognition Works", Proceedings of 2002 International conference on Image processing val. 1, 2002.
35. J. Daugman, "Complete Discrete 2D Gabor Transfors by neural Networks for Image Analysis and Comparession", Vol.36 No. 7 1988.

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

A New Method of Image Fusion Technique for Impulse Noise Removal in Digital Images

By J.Harikiran, B.Saichandana, K.Keerthi, D.Anish
GITAM University

Abstract- Image fusion is the process of combining two or more images into a single image while retaining the important features of each image. Multiple image fusion is an important technique used in military, remote sensing and medical applications. This paper presents a new method of image fusion for impulse noise removal in digital images. The images are captured by five sensors and undergo filtering by five different filtering algorithms. These five de-noised images from five different filters are combined into a single image to obtain a high quality image compared to individually de-noised image. The performance of the Image Fusion is evaluated by using a reference image quality metric, Structural similarity Index (SSIM), to estimate how well the important information in the de-noised images is represented by the fused image. Experimental results show that the fused image has more quality than other filtered images.

Keywords: *Image Fusion, Image Processing, Image Restoration, Impulse Noise*

GJCST Classification: *I.4.3*

Strictly as per the compliance and regulations of:

A New Method of Image Fusion Technique for Impulse Noise Removal in Digital Images

J.Harikiran[α], B.Saichandana[Ω], K.Keerthi[β], D.Anish[Ψ]

Abstract: Image fusion is the process of combining two or more images into a single image while retaining the important features of each image. Multiple image fusion is an important technique used in military, remote sensing and medical applications. This paper presents a new method of image fusion for impulse noise removal in digital images. The images are captured by five sensors and undergo filtering by five different filtering algorithms. These five de-noised images from five different filters are combined into a single image to obtain a high quality image compared to individually de-noised image. The performance of the Image Fusion is evaluated by using a reference image quality metric, Structural similarity Index (SSIM), to estimate how well the important information in the de-noised images is represented by the fused image. Experimental results show that the fused image has more quality than other filtered images.

Keywords: Image Fusion, Image Processing, Image Restoration, Impulse Noise.

I. Introduction

Digital images are often corrupted during acquisition, transmission or due to faulty memory locations in hardware [1]. The impulse noise can be caused by a camera due to the faulty nature of the sensor or during transmission of coded images in a noisy communication channel [2]. Consequently, some pixel intensities are altered while others remain noise free. The noise density (severity of the noise) varies depending on various factors namely reflective surfaces, atmospheric variations, noisy communication channels and so on.

In most image processing applications the images captured by different sensors are combined into a single image, which retains the important features of the images from the individual sensors, this process is known as image fusion. The images captured by multiple sensors are differently noised depending on the proximity to the object, environmental disturbances and sensor features. In this paper, the images captured by five different sensors are filtered using five different non-linear filtering algorithms such as Standard Median Filter (SMF), Component Median Filter (CMF), Vector Median Filter (VMF), Spatial Median Filter (SMF) and Modified Spatial Median Filter (MSF), producing five de-noised images. These de-noised images are fused using our fusion technique, thus obtaining a high quality image.

This paper is organized as follows, Section II presents the impulse noise in images, Section III presents five different filtering algorithms, Section IV presents Image Fusion technique, Section V presents experimental results and the paper is concluded in Section VI.

II. Impulse Noise in Images

Impulse noise [3] corruption is very common in digital images. Impulse noise is always independent and uncorrelated to the image pixels and is randomly distributed over the image. There are different types of impulse noise namely salt and pepper type of noise and random valued impulse noise. In salt and pepper type of noise the noisy pixels takes either salt value (gray level -225) or pepper value (grey level -0) and it appears as black and white spots on the images. In case of random valued impulse noise, noise can take any gray level value from zero to 225. In this case also noise is randomly distributed over the entire image and probability of occurrence of any gray level value as noise will be same.

III. Filtering Algorithms

Order-static filters are nonlinear filters whose response is based on the ordering (ranking) the pixels contained in the image area encompassed by the filter, and then replacing the value of the center pixel with the value determined by the ranking result.

The **Median Filter** [8] as the name implies, replaces the value of the pixel by the median of the intensity values in the neighborhood of that pixel defined in (1). The pixel with the median magnitude is used to replace the pixel in the signal studied.

$$MEDIANFILTER(x_1,x_2,\ldots\ldots x_N) = MEDIAN(x_1,x_2,\ldots\ldots x_N) \quad (1)$$

The median filter is more robust with respect to the presence of noise.

The **Component Median Filter** (CMF) [5], defined in (2), also relies on the statistical median concept. In the Simple Median Filter, each point in the

About[α]- Assistant Professor, Department of Information Technology, GITAM University
About[Ω]- Assistant Professor, Department of Computer Science and Engineering, GITAM University
About[β]- B.Tech Student, Department of Information Technology, GITAM University
About[Ψ]- B.Tech Student, Department of Information Technology, GITAM University
E-mail- jharikiran@gmail.com

signal is converted to a single magnitude. In the Component Median Filter, each scalar component is treated independently. A filter mask is placed over a point in the signal. For each component of each point under the mask, a single median component is determined. These components are then combined to form a new point, which is then used to represent the point in the signal studied.

$$CMF(x_1,x_2,\ldots x_N) = \begin{cases} MEDIAN(x_{1r},\ldots\ldots x_{Nr}) \\ MEDIAN(x_{1g},\ldots\ldots x_{Ng} \\ MEDIAN(x_{1b},\ldots\ldots x_{Nb}) \end{cases} \quad (2)$$

In the **Vector Median Filter** (VMF) [6] for the ordering of the vectors in a particular kernel or mask a suitable distance measure is chosen. The vector pixels in the window are ordered on the basis of the sum of the distances between each vector pixel and the other vector pixels in the window.

The sum of the distances is arranged in the ascending order and then the same ordering is associated with the vector pixels. The vector pixel with the smallest sum of distances is the vector median pixel. The vector median filter is represented as

$$X_{VMF} = \text{vectormedian (window)} \quad (3)$$

If δ_i is the sum of the distances of the i^{th} vector pixel with all the other vectors in the kernel, then

$$\delta_i = \sum_{j=1}^{N} \Delta(X_i, X_j) \quad (4)$$

where $(1 \le i \le N)$ and X_i and X_j are the vectors, N=9. $\Delta(X_i, X_j)$ is the distance measure given by the L_1 norm or the city block distance which is more suited to non correlated noise. The ordering may be illustrated as

$$\delta_1 \le \delta_2 \le \delta_3 \le, \ldots, \le \delta_9 \quad (5)$$

and this implies the same ordering to the corresponding vector pixels i.e.

$$X_{(1)} \le X_{(2)} \le, \ldots, \le X_{(9)} \quad (6)$$

where the subscripts are the ranks. Since the vector pixel with the smallest sum of distances is the vector median pixel, it will correspond to rank 1 of the ordered pixels, i.e,

$$X_{VMF} = X_{(1)} \quad (7)$$

The **Spatial Median Filter** (SMF) [5] is a uniform smoothing algorithm with the purpose of removing noise and fine points of image data while maintaining edges around larger shapes. The SMF is based on the spatial median quantile function which is a L_1 norm metric that measures the difference between two vectors. The spatial depth between a point and a set of points is defined by

$$S_{depth}(X,x_1,x_2,\ldots\ldots x_N) = 1 - \frac{1}{N-1}\left\| \sum_{i=1}^{N} \frac{X - xi}{\|X - xi\|} \right\| \quad (8)$$

Let $r_1,r_2,\ldots r_N$ represent $x_1,x_2,\ldots.x_N$ in rank order such that

$$\begin{aligned} &\ge S_{depth}(r_1,x_1,x_2,\ldots\ldots x_N) \\ &\ge S_{depth}(r_2,x_1,x_2,\ldots\ldots x_N) \\ &\ge S_{depth}(r_N,x_1,x_2,\ldots\ldots x_N) \end{aligned} \quad (9)$$

and let r_c represent the center pixel under the mask. Then

$$SMF(x_1,x_2,\ldots\ldots x_N) = r_1 \quad (10)$$

In the **Modified Spatial Median Filter** (MSMF) [5], we first calculate the spatial depth of every point within the mask and then sort these spatial depths in descending order. After the spatial depth of each point within the mask is computed, an attempt is made to use this information to first decide if the mask's center point is an uncorrupted point. If the determination is made that a point is not corrupted, then the point will not be changed. If the point is corrupted, then the point is replaced with the point with the largest spatial depth.

We can prevent some of the smoothing by looking for the position of the center point in the spatial order statistic. Let us consider a parameter P (where $1 \le P \le N$, where N represents numbers of points in the mask), which represents the estimated number of original points under a mask of points. If the position of the center mask point appears within the first P ranks of the spatial order statistic, then we can argue that while the center point is not the best representative point of the mask, it is likely to be original data and should not be replaced. The MSMF is defined by

$$MSMF(T,x_1,x_2,\ldots\ldots x_N) = \begin{cases} r_c & c \le P \\ r_1 & c > P \end{cases} \quad (11)$$

IV. Image Fusion

Given five de-noised images, it is required to combine the images into a single one that has all objects without producing details that are non-existent in the given images. Here R^1 is median filtered image, R^2 is CMF filtered image, R^3 is the VMF filtered image, R^4 is the SMF filtered image, R^5 is the MSMF filtered image. The fusion algorithm consists of the following steps:

a. Input images R^i for i=1,2,..,5 are divided into non-overlapping rectangular blocks with size of mxn (10x10 blocks). The j^{th} image blocks of R^i are referred by R^i_j .
b. Variance (*VAR*) of R^i_j are calculated for determining the sharpness values of the

corresponding blocks and the results of R^i_j are denoted by VAR^i_j. *VAR* is defined as:

$$VAR = \frac{1}{m \times n} \sum_x \sum_y (f(x,y) - \bar{f}) \quad (12)$$

Where $\bar{f}$ is the average grey level over the image.

$$\bar{f} = \frac{1}{m \times n} \sum_x \sum_y f(x,y) \quad (13)$$

c. In order to determine the sharper image block, the variances of image blocks from five images are sorted in descending order and the same ordering is associated with image blocks. The block with the maximum variance is kept in the fused image. The fusion mechanism is represented as follows:
If $VAR_{(k)}$ is the variance of block R^i_j, where k denotes the rank, the ordering of variances is given by

$$VAR_{(1)} > VAR_{(2)} > VAR_{(3)} > VAR_{(4)} > VAR_{(5)} \quad (14)$$

and this implies the same ordering to the corresponding blocks

$$R_{(1)} > R_{(2)} > R_{(3)} > R_{(4)} > R_{(5)} \quad (15)$$

Where the subscripts are the ranks of the image blocks. Since the block with the smallest variance is in the fused image, it will correspond to rank 1 of the ordered blocks ie;

$$Fused\ Block = R_{(1)} \quad (16)$$

V. Experimental Results

The proposed method of image fusion for impulse noise reduction in images was tested on the true color parrot image with 290x290 pixels. The impulse noise is added into the image with noise density 0.4. The noisy image is processed using Median, CMF, VMF, SMF and MSMF filtering algorithms. The filtered images are fused into a single image using the Image fusion method. The experimental results are shown in Figure 1. Table (1) shows the results SSIM [7] values of individual de-noised images and fused image with different noise densities.

(a) (b)

(c) (d)

(e)

(f)

(g)

(h)

Figure 1 a) original image, b) impulse noise image corrupted by noise density 0.4, c) Median Filter d) component median filter e)VMF filtered image f) SMF filtered image, g) MSMF h) Fused image

Table 1: Performance Evaluation of Fusion Method

	ND 0.2 SSIM	ND 0.4 SSIM	ND 0.6 SSIM	ND 0.8 SSIM
VMF	0.821	0.714	0.586	0.4068
SMF	0.886	0.764	0.602	0.4324
Fused Image	0.9023	0.8109	0.6434	0.4671

ND-Noise Density

VI. Conclusion

This paper presents a new method of image fusion technique for removal of impulse noise in images. The images captured by sensors undergo filtering using VMF and SMF, and then the two individual de-noised images are fused to obtain a high quality image. The Quality of the images is evaluated using Structural Similarity Index (SSIM) with different noise densities. The proposed techniques are algorithmically simple and can be used for real time imaging applications.

References Références Referencias

1. Tao Chen, Kai-Kaung Ma and Li-Hui Chen, "Tri-state median filter for image Denoising", IEEE Transactions on Image Processing, Vol 8, no.12, pp.1834-1838, December 1999.
2. Reihard Berstein," Adaptive nonlinear filters for simultaneous removal of different kinds of noise in images," IEEE Trans on circuits and systems , Vol.cas-34, no 11,pp.127-1291, Nivember 1987.
3. S.Indu, Chaveli Ramesh, " A noise fading technique for images corrupted with impulse noise", Proceedings of ICCTA07, IEEE.

4. F.VanderHeijden, Image Based Measurement Systems. Newyork, Wiley, 1994.
5. James C. Church, Yixin Chen, and Stephen V. Rice , "A Spatial Median Filter for Noise Removal in Digital Images", 2008 IEEE.
6. R. H. Laskar, B. Bhowmick, R. Biswas and S. Kar ," Removal of impulse noise from color image" , 2009 IEEE.
7. Z. Wang and A. C. Bovik, "A universal image quality index," IEEE Signal Processing Letters, vol. 9, pp. 81–84, Mar. 2002.
8. K. Jain, Fundamentals of Digital Image Processing. Englewood Cliffs, NJ: Prentice-Hall, 1989.

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

Improve Speech Enhancement Using Weiner Filtering

By S.China Venkateswarlu, Dr. K.Satya Prasad, Dr. A.SubbaRami Reddy

Abstract- Speech enhancement aims to improve speech quality by using various algorithms. It may sound simple, but what is meant by the word quality. It can be at least clarity and intelligibility, pleasantness, or compatibility with some other method in speech processing. Wiener filter are rather simple and workable, but after the estimation of the background noise, one neglects the fact that the signal is actually speech. Furthermore, the phase component of the signal is left untouched. However, this is perhaps not such a bad problem; after all, human ear is not very sensitive to phase changes. The third restriction in spectral subtraction methods is the processing of the speech signal in frames, so the Proceeding from one frame to another must be handled with care to avoid discontinuities. Noise reduction is a key-point of speech enhancement systems in hands-free communications. A number of techniques have been already developed in the frequency domain such as an optimal short-time spectral amplitude estimator proposed by Ephraim and Malah including the estimation of the a priori signal-to-noise ratio. This approach reduces significantly the disturbing noise and provides enhanced speech with colorless residual noise. In this paper, we propose a technique based on a Wiener filtering under uncertainty of signal presence in the noisy observation. Two different estimators of the a priori signal-to-noise ratio are tested and compared. The main interest of this approach comes from its low complexity. In this paper we demonstrate the application of weiner filter for a speech signal using Matlab 7.1 and signal processing toolbox.

Keywords: *Communication, Enhancement, Intelligibility, Matlab, Speech, Wiener filter*

GJCST Classification: *H.5.2, I.2.7*

Strictly as per the compliance and regulations of:

Improve Speech Enhancement Using Weiner Filtering

S.China Venkateswarlu[α], Dr. K.Satya Prasad[Ω], Dr. A.SubbaRami Reddy[β]

Abstract- Speech enhancement aims to improve speech quality by using various algorithms. It may sound simple, but what is meant by the word quality. It can be at least clarity and intelligibility, pleasantness, or compatibility with some other method in speech processing. Wiener filter are rather simple and workable, but after the estimation of the background noise, one neglects the fact that the signal is actually speech. Furthermore, the phase component of the signal is left untouched. However, this is perhaps not such a bad problem; after all, human ear is not very sensitive to phase changes. The third restriction in spectral subtraction methods is the processing of the speech signal in frames, so the Proceeding from one frame to another must be handled with care to avoid discontinuities. Noise reduction is a key-point of speech enhancement systems in hands-free communications. A number of techniques have been already developed in the frequency domain such as an optimal short-time spectral amplitude estimator proposed by Ephraim and Malah including the estimation of the a priori signal-to-noise ratio. This approach reduces significantly the disturbing noise and provides enhanced speech with colorless residual noise. In this paper, we propose a technique based on a Wiener filtering under uncertainty of signal presence in the noisy observation. Two different estimators of the a priori signal-to-noise ratio are tested and compared. The main interest of this approach comes from its low complexity. In this paper we demonstrate the application of weiner filter for a speech signal using Matlab 7.1 and signal processing toolbox.

Keywords- Communication, Enhancement, Intelligibility, Matlab, Speech, Wiener filter

I. Introduction

Speech is a form of communication in every day life. It existed since human civilizations began and even till now, speech is applied to high technological telecommunication systems. A particular field which I personally feel will excel will be speech signal processing in the world of telecommunications. As applications like cellular and satellite technology are getting popular among mankind, human beings tend to demand more advance technology and are in search of improved applications. For this reason, researchers are looking closely into the four generic attributes of speech coding. They are complexity, quality, bit rate and delay. Other issues like robustness to transmission errors, multistage encoding/decoding, and accommodation of non-voice signals such as in-band signaling and voice band modem data play an important role in coding of speech as well .

In order to understand these processes, both human and machine speeches have to be studied carefully on the structures and functions of spoken language:

How we produce and perceive it and how speech technology may assist us in communication. Therefore in this project, we will be looking more into speech processing with the aid of an interesting technology known as the Weiner Filter for speech processing. Presently, this technique is widely used in the field of signal processing. Speech processing has been a growing and dynamic field for more than two decades and there is every indication that this growth will continue and even accelerate. During this growth there has been a close relationship between the development of new algorithms and theoretical results, new filtering techniques are also of consideration to the success of speech processing. One of the common adaptive filtering techniques that are applied to speech is the Wiener filter. This filter is capable of estimating errors however at only very slow computations. On the other hand, the Kalman filter suppresses this disadvantage.

As widely known to the world, weiner filtering techniques are used on GPS (Global Positioning System) and INS (Inertial Navigation System). They are widely used for speech signal coding applications. Due to its accurate estimation characteristic, electrical engineers are picturing the Weiner filter as a design tool for speech, whereby it can estimate and resolve errors that are contained in speech after passing through a distorted channel. Due to this motivating fact, there are many ways a Weiner filter can be tuned to suit engineering applications such as network telephony and even satellite phone conferencing. Knowing the fact

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

About[α]- HITS College of Engineering is a RESEARCH SCHOLAR, from JNTU Kakinada, A.P., INDIA., RESEARCH SCHOLAR ROLL NO: 09022P0426, phone: 08415-242500, fax: 08415-24500252, Mobile No: +91 9666819730, +91 9030076793;
E-mail: cvenkateswarlus@gmail.com
About[Ω]- Professor & DE from JNTU Kakinada, Kakinada.and Research Supervisor from JNTUCollege of Engineering Kakinada..
E-mail: prasad_kodati@yahoo.co.in.
About[β]- Professor & Principal from SKIT-Chittor.and Research Supervisor from JNTUK.
E mail: principalskit@gmail.com

that preserving information, which is contained in speech, is of extreme importance, the availability of signal filters such as the Weiner filter is of great importance.

In this paper, the primary goal is to design a MATLAB based simulator for processing of speech together with the aid of the Weiner filtering technique and to obtain a reconstructed speech signal, which is similar to the input speech signal. To achieve these results, sample speeches were obtained. These were modeled as an autoregressive (AR) process and represented in the state-space domain by the Weiner filter. The original speech signal and the reconstructed speech signal obtained from the output of the filter were compared. The idea of this comparison is to pursue an output speech signal, which is similar to the original one. It was concluded that Weiner filtering is a good constructing method for speech.

II. Speech Processing

The term speech processing basically refers to the scientific discipline concerning the analysis and processing of speech signals in order to achieve the best benefit in various practical scenarios. The field of speech processing is, at present, undergoing a rapid growth in terms of both performance and applications. This is stimulated by the advances being made in the field of microelectronics, computation and algorithm design. Nevertheless, speech processing still covers an extremely broad area, which relates to the following three engineering applications:

- Speech Coding and transmission that is mainly concerned with man-to man voice Communication
- Speech Synthesis which deals with machine-to-man communications;
- Speech Recognition relating to man-to machine communication.

a) *Speech Coding*

Speech coding or compression is the field concerned with compact digital representations of speech signals for the purpose of efficient transmission or storage. The central objective is to represent a signal with a minimum number of bits while maintaining perceptual quality. Current applications for speech and audio coding algorithms include cellular and personal communications networks (PCNs), teleconferencing, desktop multi-media systems, and secure communications.

b) *Speech Synthesis*

The process that involves the conversion of a command sequence or input text (words or sentences) into speech waveform using algorithms and previously coded speech data is known as speech synthesis. The inputting of text can be processed through by keyboard, optical character recognition, or from a previously stored database. A speech synthesizer can be characterized by the size of the speech units they concatenate to yield the output speech as well as by the method used to code, store and synthesize the speech. If large speech units are involved, such as phrases and sentences, high-quality output speech (with large memory requirements) can be achieved. On the contrary, efficient coding methods can be used for reducing memory needs, but these usually degrade speech quality.

c) *Speech Recognition*

Speech or voice recognition is the ability of a machine or program to recognize and carry out voice commands or take dictation. On the whole, speech recognition involves the ability to match a voice pattern against a provided or acquired vocabulary. A limited vocabulary is mostly provided with a product and the user can record additional words. On the other hand, sophisticated software has the ability to accept natural speech (meaning speech as we usually speak it rather than carefully-spoken speech).

Speech information can be observed and processed only in the form of sound waveforms. It is an essential for speech signal to be reconstructed properly.

d) *Speech Production*

Speech is the acoustic product of voluntary and well-controlled movement of a vocal mechanism of a human. During the generation of speech, air is inhaled into the human lungs by expanding the rib cage and drawing it in via the nasal cavity, velum and trachea it is then expelled back into the air by contracting the rib cage and increasing the lung pressure. During the expulsion of air, the air travels from the lungs and passes through vocal cords which are the two symmetric pieces of ligaments and muscles located in the larynx on the trachea. Speech is produced by the vibration of the vocal cords. Before the expulsion of air, the larynx is initially closed. When the pressure produced by the expelled air is sufficient, the vocal cords are pushed apart, allowing air to pass through. The vocal cords close upon the decrease in air flow. This relaxation cycle is repeated with generation frequencies in the range of 80Hz – 300Hz. The generation of this frequency depends on the speaker's age, sex, stress and emotions. This succession of the glottis openings and closure generates quasi-periodic pulses of air after the vocal cords.

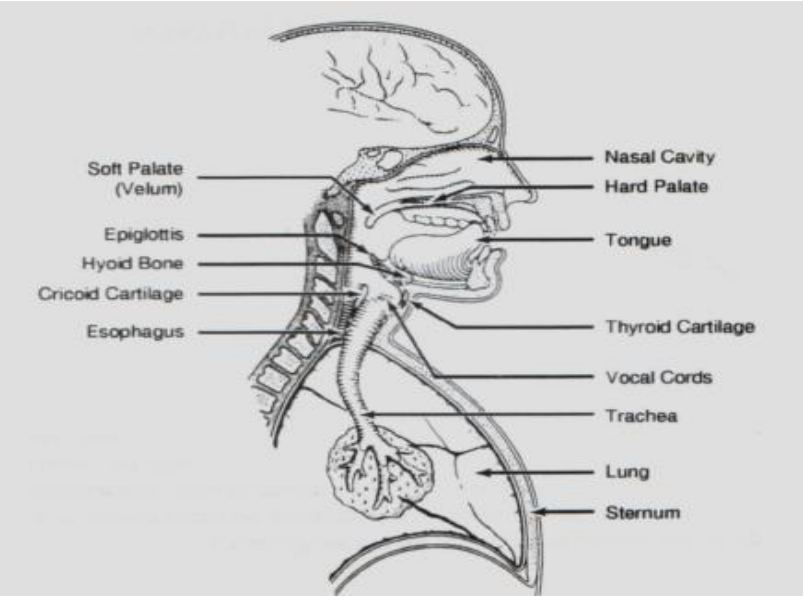


Figure1: Speech -acoustic product of voluntary and well controlled movement of a vocal mechanism of a human

The speech signal is a time varying signal whose signal characteristics represent the different speech sounds produced. There are three ways of labelling events in speech. First is the silence state in which no speech is produced. Second state is the unvoiced state in which the vocal cords are not vibrating, thus the output speech waveform is aperiodic and random in nature. The last state is the voiced state in which the vocal cords are vibrating periodically when air is expelled from the lungs. This results in the output speech being quasi-periodic- shows a speech waveform with unvoiced and voiced state.

Speech is produced as a sequence of sounds. The type of sound produced depends on shape of the vocal tract. The vocal tract starts from the opening of the vocal cords to the end of the lips. Its cross sectional area depends on the position of the tongue, lips, jaw and velum. Therefore the tongue, lips, jaw and velum play an important part in the production of speech.

III. Hearing and Perception

Audible sounds are transmitted to the human ears through the vibration of the particles in the air. Human ears consist of three parts, the outer ear, the middle ear and the inner ear. The function of the outer ear is to direct speech pressure variations toward the eardrum where the middle ear converts the pressure variations into mechanical motion. The mechanical motion is then transmitted to the inner ear, which transforms these motion into electrical potentials that passes through the auditory nerve, cortex and then to the brain . Figure below shows the schematic diagram of the human ear.

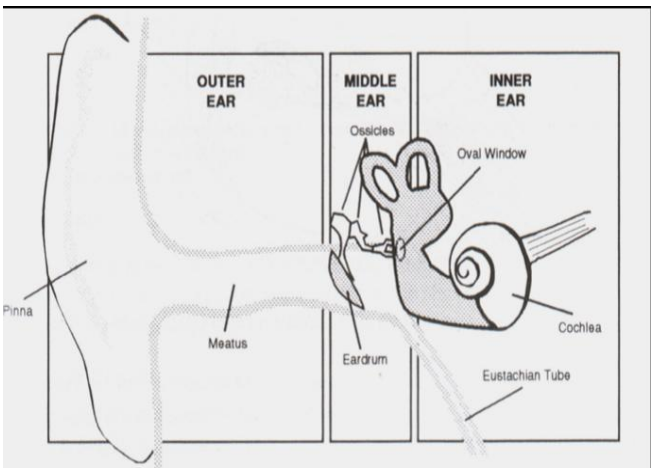


Figure2: The schematic diagram of the human ear

IV. Speech Waveform

The speech waveform needs to be converted into digital format before it is suitable for processing in the speech recognition system. The raw speech waveform is in the analog format before conversion. The conversion of analog signal to digital signal involves three phases, mainly the sampling, quantisation and coding phase. In the sampling phase, the analog signal is being transformed from a waveform that is continuous in time to a discrete signal. A discrete signal refers to the sequence of samples that are discrete in time. In the quantisation phase, an approximate sampled value of a variable is converted into one of the finite values contained in a code set. These two stages allow the speech waveform to be represented by a sequence of values with each of these values belonging to the set of finites values. After passing through the sampling and quantization stage, the signal is then coded in the

coding phase. The signal is usually represented by binary code. These three phases needs to be carried out with caution as any miscalculations, over-sampling and quantization noise will result in loss of information. Below are the problems faced by the three phases.

V. Single and Multi-Channel Enhancement

Single channel methods operate on the input obtained from only one microphone. They have been attractive due to cost and size factors, especially in mobile communications. In contrast, multi-channel methods employ an array of two or more microphones to record the noisy signal and exploit the resulting spatial diversity. The two approaches are not necessarily independent, and can be combined to improve performance. For example, in practical diffuse noise environments, the multi-channel enhancement schemes rely on a single-channel post-filter to provide additional noise reduction. We discuss single-channel methods and introduce the contributions of this project towards this area is also included in this document. This section is intended to be a survey on single-channel enhancement algorithms.

VI. Maximum-Likelihood Estimation

Consider the estimation of a parameter $\mu = [\mu 1 : : : \mu p]^T$ based on a sequence of K observations $y = [y(0) : : : y(Ki1)]^T$. In ML estimation, μ is treated as a deterministic variable. The ML estimate of μ is the value μ ML that maximizes the likelihood function p(y; μ) defined on the data. ML estimation has several favorable properties, in particular, it is asymptotically unbiased and efficient, i.e., as the number of observations K tends to infinity, the ML estimate is unbiased and achieves the Cramer-Rao lower bound (CRLB). It can be shown that

$$\theta^{\mathrm{ML}} \underset{K\to\infty}{\sim} \mathcal{N}(\theta, \mathbf{I}^{-1}(\theta)), \tag{1}$$

where I(μ) is the p £ p Fisher information matrix whose (i; j)th entry is given by

$$[\mathbf{I}(\theta)]_{ij} = -\mathrm{E}\left[\frac{\partial^2 \ln p(\mathbf{y};\theta)}{\partial\theta_i \partial\theta_j}\right] \tag{2}$$

Thus we have (asymptotically)

$$\begin{aligned} &\text{Unbiased: } \mathrm{E}[\theta^{\mathrm{ML}}] = \theta, \\ &\text{CRLB: } \mathrm{var}(\theta_i^{\mathrm{ML}}) = [\mathbf{I}^{-1}(\theta)]_{ii}. \end{aligned} \tag{3}$$

The maximization of the likelihood function is performed over the domain of μ. In many cases, μ ML cannot be computed in closed form and a numerical solution is obtained instead. Such numerical solutions are typically obtained through iterative maximization procedures such as the Newton-Raphson method or the expectation-maximization (EM) approach. The initial value of the parameter used to start the iterative procedure usually has a strong impact on whether the final estimate results in a local or a global maximum of the likelihood function.

In applications where the parameter μ is known to assume one of a finite set of values, the problems due to the iterative procedures can be avoided by performing the maximization over this finite set. An exhaustive search over the finite parameter space guarantees a global maximum. For speech enhancement, we assume that both speech and noise can be described by independent auto-regressive (AR) processes. The problem is then one of estimating the speech and noise LP coefficients based on the observed noisy speech in an ML framework. The clean speech AR model can be mathematically expressed as

$$x(n) = \sum_{l=1}^{p} a_l x(n-l) + e(n), \tag{4}$$

where a1…. ap are the LP coefficients of order p and e(n) is the prediction error, also referred to as the excitation signal. It is common to model e(n) as a Gaussian random process. The LP analysis is typically performed for each frame of 20-30 ms, within which speech can be assumed to be stationary.

For each frame, the model parameters are the vector of LP coefficients $\mu = [a1 : : : ap]$, and the variance of the excitation signal. A similar model can be obtained for the noise signal. The physiology of speech production imposes a constraint on the possible shapes of the speech spectral envelope. Since the spectral envelope is specified by the LP coefficients, this knowledge can be modeled using a sufficiently large codebook of speech LP coefficients obtained from long sequences of training data. Such a-priori information about the LP coefficients of speech has been exploited successfully in speech coding using trained codebooks.

Similarly, noise LP coefficients can also be modeled based on training sequences for different noise types. Thus, it is sufficient to perform the maximization over the speech and noise codebooks.

We characterize the speech and noise power spectra, which can be used to construct a Wiener filter to obtain the enhanced speech signal. Given the noisy data, the excitation variances maximizing the likelihood are determined for each pair of speech and noise LP coefficients from the codebooks. This is done for all combinations of codebook pairs, and the most likely codebook combination, together with the optimal excitation variances, is obtained. Since this optimization

is performed on a frame-by-frame basis, good performance is achieved in non-stationary noise environments.

Apart from restricting the search space, using a codebook in the ML estimation has an additional benefit in applications where a codebook index needs to be transmitted over a network, e.g., in speech coding. In this case, the likelihood function can be interpreted as a modified distortion criterion.

VII. Bayesian Mmse Estimation

In ML estimation, the parameter μ is treated as a deterministic but unknown constant. In the Bayesian approach, μ is treated as a random variable. The Bayesian methodology allows us to incorporate prior (before observing the data) knowledge about the parameter by assigning a prior pdf to μ.

A cost function is formulated and its expected value, referred to as the Bayesian risk, is minimized. A commonly used cost function is the mean squared error (MSE).

In this case, the Bayesian minimum mean squared error (MMSE) estimate μ^{BY} of μ given the observations y is obtained by minimizing E[(μ i μ^{BY})2], where E is the statistical expectation operator. The expectation is with respect to the joint distribution p(y; μ). Thus, the cost function to be minimized can be written as

$$\begin{aligned}\eta &= \mathrm{E}[(\theta-\theta^{\mathrm{BY}})^2] \\ &= \int\int(\theta-\theta^{\mathrm{BY}})^2 p(\mathbf{y},\theta)d\mathbf{y}d\theta \\ &= \int\left(\int(\theta-\theta^{\mathrm{BY}})^2 p(\theta|\mathbf{y})d\theta\right)p(\mathbf{y})d\mathbf{y},\end{aligned} \tag{5}$$

where the posterior pdf p(μjy) is the pdf of μ after the observation of data. Since p(y) ¸ 0, it is su±cient to minimize the inner integral for each y. An estimate of μ can be found by determining a stationary point of the cost function (setting the derivative of the inner integral to zero). We can write

$$\frac{\partial}{\partial\theta^{\mathrm{BY}}}\int(\theta-\theta^{\mathrm{BY}})^2 p(\theta|\mathbf{y})d\theta = 0 \tag{6}$$

$$\theta^{\mathrm{BY}} = \int\theta p(\theta|\mathbf{y})d\theta = \mathrm{E}[\theta|\mathbf{y}]. \tag{7}$$

E[μjY = y], where y is a realization of the corresponding random variable Y. Using Bayes' rule, the posterior pdf can be written as

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}, \tag{8}$$

where the denominator p(y) is a normalizing factor, independent of the parameter μ.

We describe a method to obtain Bayesian MMSE estimates of the speech and noise AR parameters. The respective prior pdfs are modeled by codebooks. The integral in (7) is replaced by a summation over the codebook entries. We also consider MMSE estimation of functions of the AR parameters, and one such function is shown to result in the MMSE estimate of the clean speech signal, given the noisy speech. As in the ML case, MMSE estimates of the speech and noise AR parameters are obtained on a frame-by-frame basis, ensuring good performance in non stationary noise.

In the ML estimation framework, one pair of speech and noise codebook vectors was selected as the ML estimate, whereas the Bayesian approach results in a weighted sum of the speech (noise) codebook vectors. The Bayesian method provides a framework to account for both the knowledge provided by the observed data and the prior knowledge.

VIII. Single-Channel Speech Enhancement

Single-channel speech enhancement systems obtain the input signal using only one microphone. This is in contrast to multi-channel systems where the presence of two or more microphones enables both spatial and temporal processing. Single-channel approaches are relevant due to cost and size factors. They achieve noise reduction by exploiting the spectral diversity between the speech and noise signals. Since the frequency spectra of speech and noise often overlap, single-channel methods generally achieve noise reduction at the expense of speech distortion.

The reduction of background noise using single-channel methods requires an estimate of the noise statistics. Early approaches were based on voice activity detectors (VAD), and noise estimates were updated during periods of speech inactivity. Accuracy deteriorates with decreasing signal-to-noise ratios (SNR) and in non-stationary noise. Soft-decision VADs, update the noise statistics even during speech activity.

Since single-channel methods exploit the spectral diversity between the speech and noise signals, it is therefore natural to perform the processing in the frequency domain. Processing is done on short segments of the speech signal, typically of the order of 20 to 30 ms, to ensure that the speech signal satisfies assumptions of wide-sense stationary. The

segmentation is performed using a sliding window of finite support. The windowed signal (assuming it is absolute sum able) is transformed to the frequency domain using the discrete short-time Fourier transform (STFT)

$$X_m(k) = \frac{1}{\sqrt{K}} \sum_{n=-\infty}^{\infty} x(n)h(n-m)\exp(-j\frac{2\pi}{K}kn),\ k = 0,1,\ldots,K-1, \quad (9) \qquad (9)$$

1. Math and computation
2. Algorithm development
3. Data acquisition
4. Data analysis ,exploration ands visualization
5. Scientific and engineering graphics

MATLAB displays graphs in a special window known as a figure. To create a graph, you need to define a coordinate system. Therefore every graph is placed within axes, which are contained by the figure. The actual visual representation of the data is achieved with graphics objects like lines and surfaces. These objects are drawn within the coordinate system defined by the axes, which MATLAB automatically creates specifically to accommodate the range of the data. The actual data is stored as properties of the graphics objects.

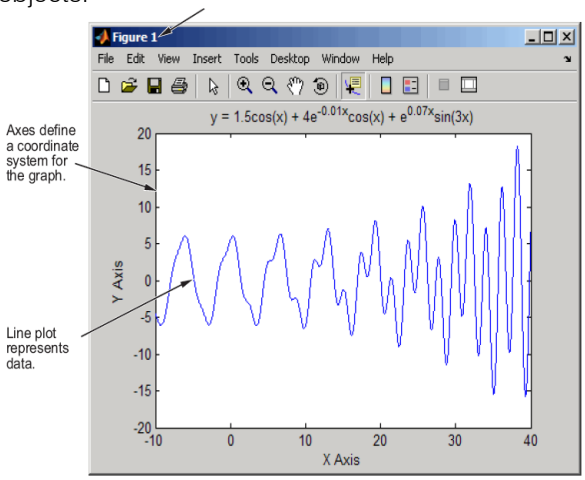


Figure3: Graphical objects

a) *Plotting Tools*

Plotting tools are attached to figures and create an environment for creating Graphs. These tools enable you to do the following:

Select from a wide variety of graph types , Change the type of graph that represents a variable , See and set the properties of graphics objects , Annotate graphs with text, arrows, etc.

• Create and arrange subplots in the figure ,• Drag and drop data into graphs Display the plotting tools from the View menu or by clicking the plotting tools icon in the figure toolbar, as shown in the following picture

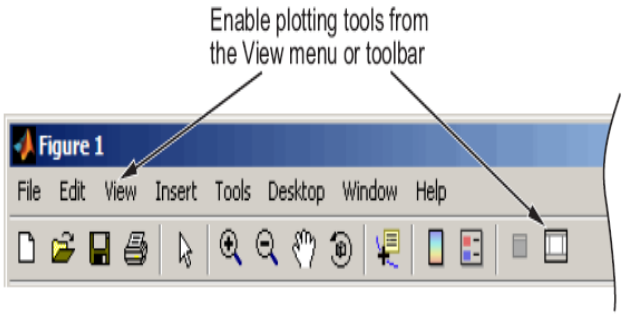


Figure4: Drag and drop data into graphs display

b) *Editor/Debugger*

Use the Editor/Debugger to create and debug M-files, which are programs you write to run MATLAB functions. The Editor/Debugger provides a graphical user interface for text editing, as well as for M-file debugging. To create or edit an M-file use File > New or File > Open, or use the edit function.

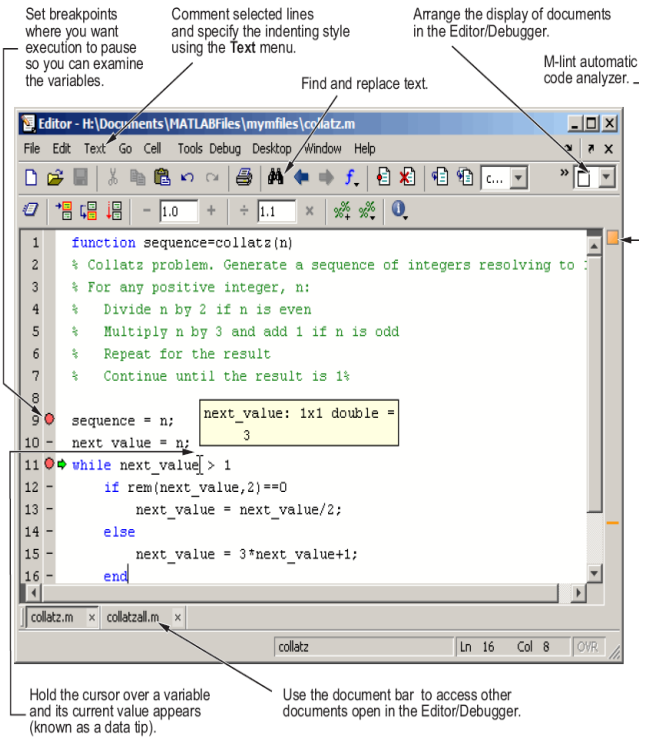


Figure5: M-file debugging-edit function

IX. Results

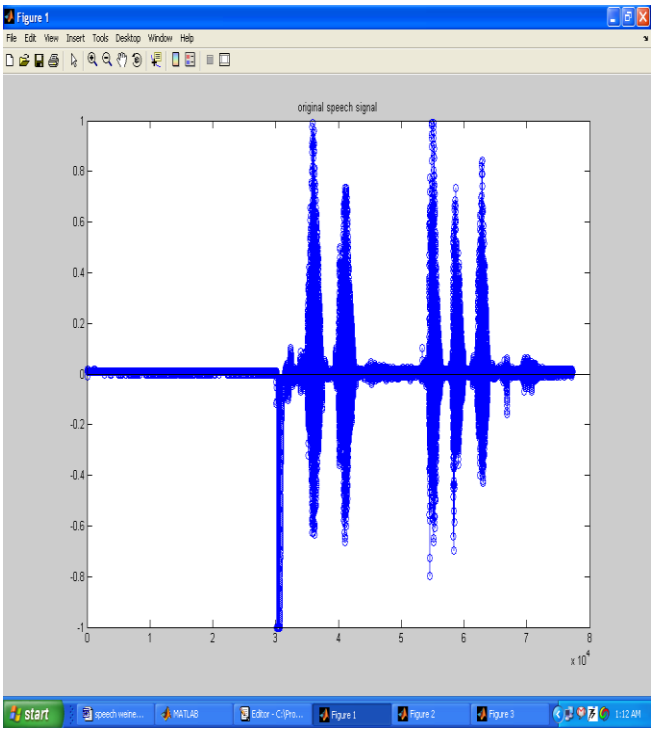


Figure6: Original speech signal

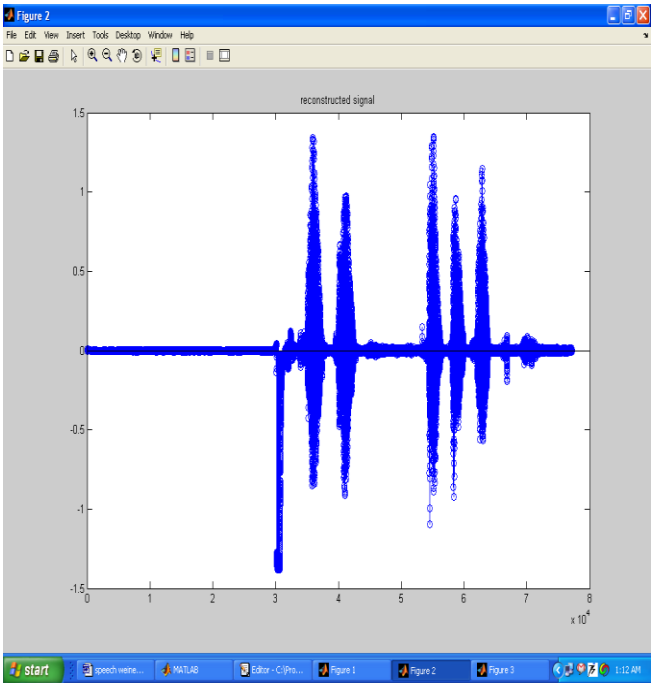


Figure6: Reconstructed signal

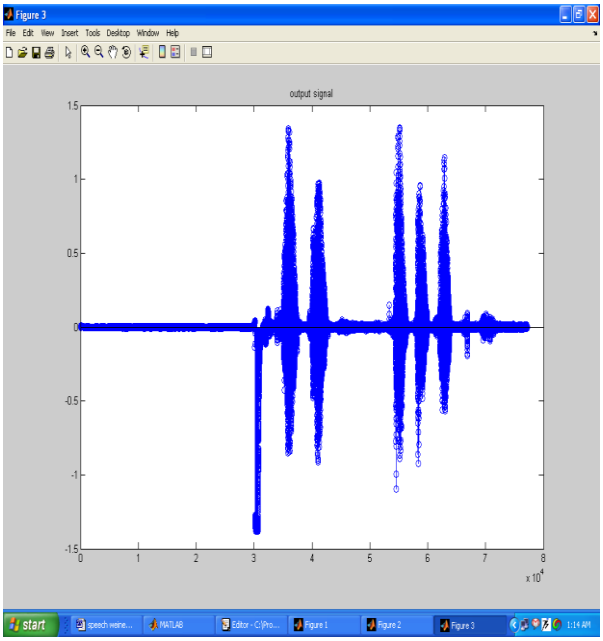


Figure7: Out put signal

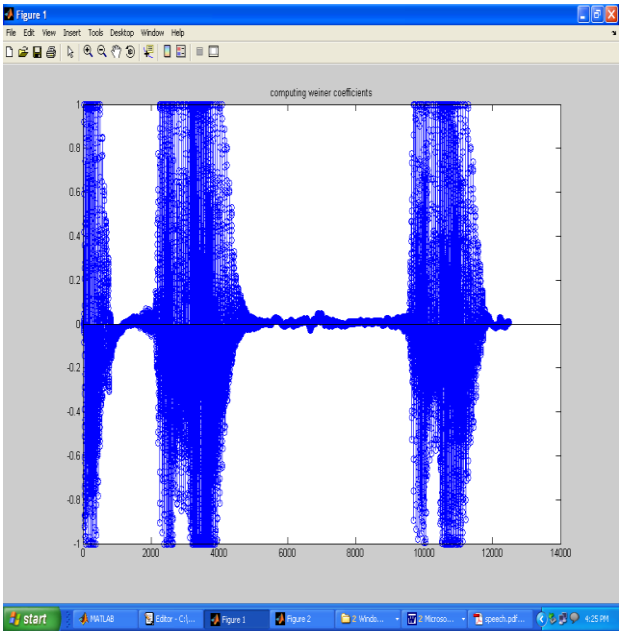


Figure8: Computed Weiner coefficients

Total no.of coefficients 100 The shift length have to be an integer as it is the number of samples. shift length is fixed to 80.

X. Conclusion

In this paper, an implementation of employing Weiner filtering to speech processing had been developed. As has been previously mentioned, the purpose of this approach is to reconstruct an output speech signal by making use of the accurate estimating ability of the Weiner filter. True enough, simulated results from the previous chapter had proven that the

Kalman filter indeed has the ability to estimate accurately. Furthermore, the results have also shown that Weiner filter could be tuned to provide optimal performance. This test is of necessity for the reason that different signals are bound to be similar but not identical. By and large, this thesis has been quite successful in terms of achieving the objectives. Consequently, perception on signal processing and Kalman filter had also been treasured throughout the process. Most importantly, the skill in time management applied during the research of this project had been developed.

XI. Future Developments

The future of Weiner filtering on Speech Processing seems reasonably bright. During the process of this project, many issues have been found to be potential topics for further research work. For that reason, the following issues were raised for further developments:

Speech Compression: The technique of Kalman filtering can be applied to speech coding using Autoregressive (AR) modeling. Since compression is the major issue here, optimal compression cannot be achieve if the entire speech signal used. The best approach is to extract the excitation sequence (white noise) otherwise known as the nonredundant information, which contains the core information of the entire speech signal. Moreover, it is said to beneficial for compression. After which, this white noise can be process through a quantizer and ready to be encoded into suitable bit rates. Quality of speech: Human speech is however difficult to artificially produce. This implies that the quality of speech is yet another major issue. Quality factors to be considering which will affect the speech are somehow complex. For instance, tandem connections, robustness to acoustic inputs, robustness to digital transmission errors as well as delay of transmission are all important factors for thorough investigation.

References Références Referencias

1. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol.27, pp. 113-120, Apr. 1979.
2. M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. *IEEE Int. Conf. On Acoust. Speech, Signal Procs.*, pp. 208-211, Apr. 1979.
3. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-term spectral amplitude estimator," *IEEE Trans.On Acoust. Speech, Signal Proc.,* vol.ASSP-32, No.6, pp. 1109-1121, Dec.1984.
4. Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," IEEE *Trans. Speech Audio Procs,* vol. 3, pp. 251-266, Jul. 1995.
5. J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE,* vol. 67, No. 12, pp. 221-239, Dec. 1979.
6. M. Sambur, "Adaptive noise canceling for speech signals," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 26, p. 419-423, 1978.
7. S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing Speech corrupted by colored noise," Proc ICASSP, pp.IV-4164, May 2002.
8. Alan V. Oppenheim and Ronald W. Schafer. "Discrete-Time Signal Processing".
9. Lawrence R.Rabiner, Ronald W.Schafer. "Digital Processing of Speech Signals". Pearson education

XII. Appendix

M-files for MULTIBAND SPECTRAL SUBTRACTION
Main module

```
%*********************************************
**********************Real is a file containing the words "PHASE DETECTION AND RECOGNITION ARE CHALLENGING TASKS" uttered by a male voice.
%*********************************************
**********************
 [signal,fs]                                    =
wavread('C:\MATLAB7\work\noisy\real.wav');
% signal = signal(1:45000);
nsignal = signal;
% EACH BAND HAS N/bands FREQUNCY COMPONENTS
each = N/bands;
[seg, nf] = segmenth(ensignal,ovlap,W);
% DETERMINE THE DFT OF NOISY SPEECH
pha = (angle(fft(seg,N,2)));
% ESTIMATE OF NOISE FROM FIRST 10 FRAMES
nmag = (abs(fft(seg(1:10,:),N,2)));
uw = (sum(nmag))/10;
% MAGNITUDE AVERAGING
```

XIII. Acknowledgment

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

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

Web Page Prediction for Web Personaliation:A Review

By R. Khanchana, Dr. M. Punithavalli
Karpagam University

Abstract- This paper proposes a survey of Web Page Ranking for web personalization. Web page prefetching has been widely used to reduce the access latency problem of the Internet. However, if most prefetched web pages are not visited by the users in their subsequent accesses, the limited network bandwidth and server resources will not be used efficiently and may worsen the access delay problem. Therefore, it is critical that we have an accurate prediction method during prefetching. The technique like Markov models have been widely used to represent and analyze user‘s navigational behavior (usage data) in the Web graph, using the transitional probabilities between web pages, as recorded in the web logs. The recorded users‘ navigation is used to extract popular web paths and predict current users‘ next steps.

Keywords: *Web Personalization, Page Ranking, User Browsing, Markov Model.*

GJCST Classification: *H.3.5*

Strictly as per the compliance and regulations of:

Web Page Prediction for Web Personalization: A Review

R. Khanchana[α], Dr. M. Punithavalli[Ω]

Abstract- This paper proposes a survey of Web Page Ranking for web personalization. Web page prefetching has been widely used to reduce the access latency problem of the Internet. However, if most prefetched web pages are not visited by the users in their subsequent accesses, the limited network bandwidth and server resources will not be used efficiently and may worsen the access delay problem. Therefore, it is critical that we have an accurate prediction method during prefetching. The technique like Markov models have been widely used to represent and analyze user's navigational behavior (usage data) in the Web graph, using the transitional probabilities between web pages, as recorded in the web logs. The recorded users' navigation is used to extract popular web paths and predict current users' next steps.

Keywords- Web Personalization, Page Ranking, User Browsing, Markov Model.

I. Introduction

A portion of Data mining, which resolves around the assessment of the World Wide Web, is known as web mining. Data mining, Internet Technology, World Wide Web as well as semantic web, are incorporated in web mining. Web mining refers to the use of data mining techniques to automatically discover and extract information form world wide web documents and services. Web mining has been classified into three areas such as Web content mining, Web structure mining and Web usage mining. The most common applications include the ranking of the results of a web search engine and the provision of recommendations to users of (usually commercial) web sites, known as web personalization. Even with the speed of today's Internet, web latency is still one of the major concerns of its users. Reducing latency is particularly important for online businesses, since if their web pages cannot be opened within about eight seconds, they might lose customers. Web servers collect huge amount of data everyday. Users search any information, that relevant data is prefetched from web server.

About[α]- Research Scholar, Karpagam University, Coimbatore, Tamilnadu,India-641 021.
E-mail : kanchusri@gmail.com
About[Ω]- Director and Head, SNS college, Bharathiar University, Coimbatore, Tamilnadu, India.
E-mail : mpunitha_srcw@yahoo.co.in

However, if most prefetched web pages are not visited by the user in their subsequent accesses, the limited network bandwidth and server resources will not be used efficiently and may worsen the access latency problem. The objective of a Web personalization system is to "provide users with the information they want or need, without expecting from them to ask for it explicitly" [14]. The most common approaches used for web user browsing pattern prediction are Markov model, sequential association rules and clustering. PageRank is used in order to rank web pages based on the results returned by a search engine after a user query. The ranking is performed by evaluating the *importance* of a page in terms of its connectivity to and from other important pages. In the context of navigating a web site, a page/path is *important* if many users have visited it before, we propose a novel approach that is based on a personalized version of PageRank, applied to the navigational tree created by the previous users' navigations. A new technique was proposed by Page and Brin called PageRank to compute the importance of Web pages. PageRank [2] determines the significance of Web pages and helps a search engine to choose high quality pages more efficiently.

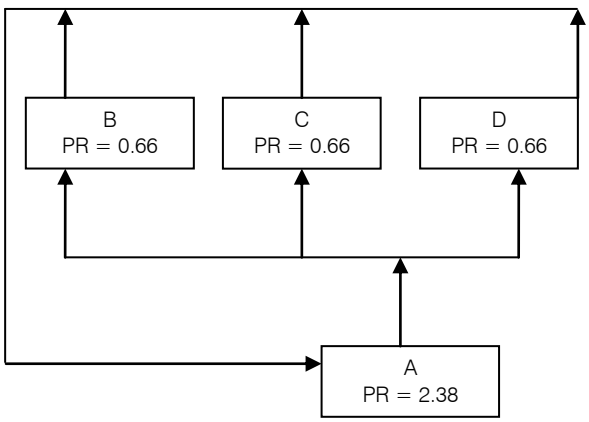


Fig.1 Web Page Linking and their Page Ranks

In the above example we consider the average page rank for the site is 1. The page B, C and D has the page ranks like 0.66 but A has 2.38.The links are well done such that we can navigate in and out of every page. The PR has also been distributed in favour of page A which has PR 2.38. This knowledge is then used

from the system in order to personalize the site according to each user's behavior and profile.

Fig 2 shows a multi-tiered Web site and the caching and personalization techniques suitable for each Web site component. The caching levels show that performance is maximized when cache hits occur close to the browsers. For example, at the ISP and router levels, rule-based and simple filtering may offer sufficient personalization capabilities for a relatively small investment of effort. When more is needed or wanted, more complex techniques can be implemented. When data mining is needed to develop business intelligence and offer highly sophisticated personalization, the processing occurs at the database layer.

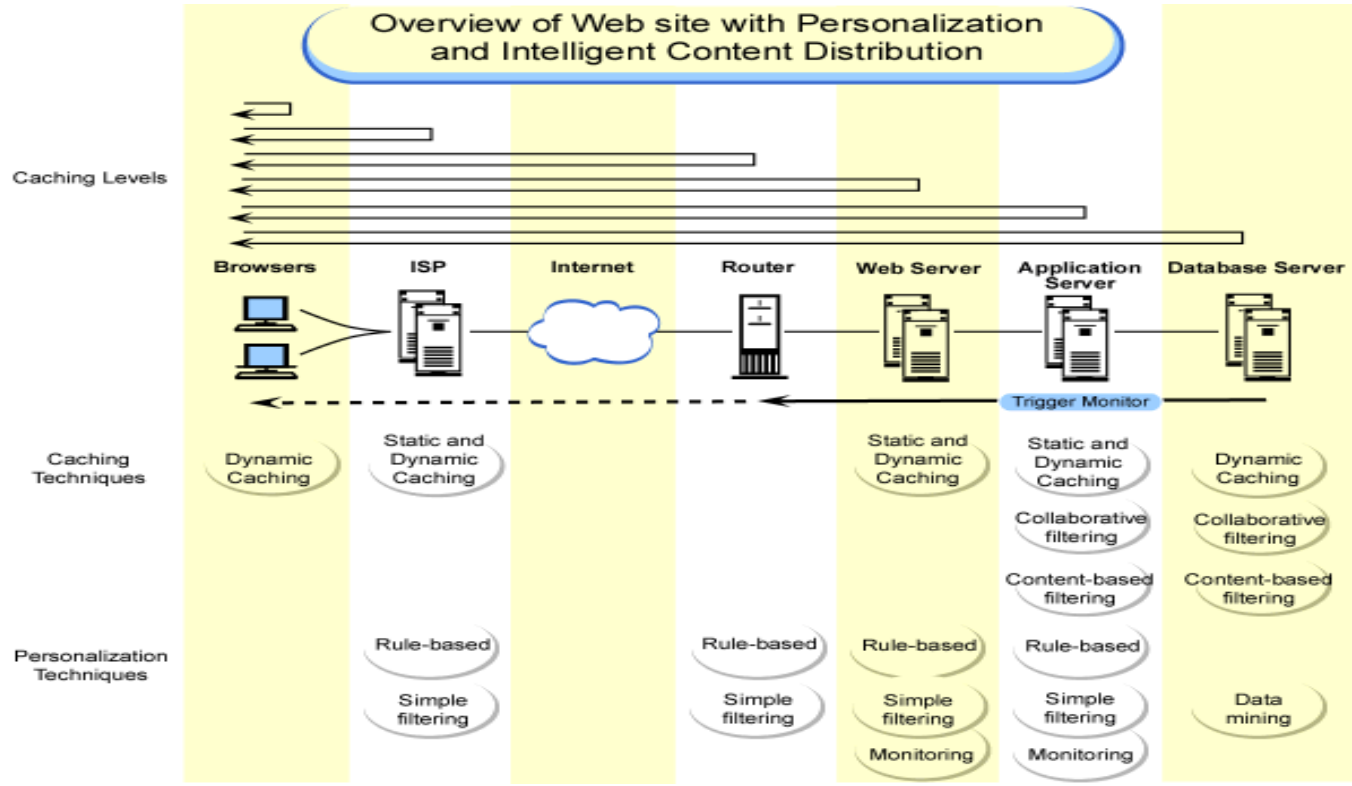


Fig 2. Overview of Web site with personalization and intelligent content distribution

The basic personalization techniques are

1. Rule Based
2. Simple Filtering
3. Content- Based Filtering
4. Collaborative Filtering

Rule Based: Rule-based techniques provide a visual editing environment for the business administrator to specify business rules to drive personalization.

Simple Filtering: Simple filtering relies on predefined groups, or classes, of visitors to determine what content is displayed or what service is provided.

Content- Based Filtering: Content-based filtering works by analyzing the content of the objects to form a representation of the visitor's interests.

Collaborative Filtering: Collaborative filtering collects visitors' opinions on a set of objects, using either explicit or implicit ratings, to predict a particular visitor's interest in an item.

Caching techniques have long been used to improve the system performance. With content caching, frequently accessed pages do not need to be retrieved remotely or materialized at the server for each access. This can significantly reduce the latency for obtaining Web pages, as well as reduce the load on the server and network. In the Web environment, frequently accessed Web pages can be cached at the client browser, proxy servers, and server caches. For caching to be effective, data needs to be reused frequently. Personalization is a process of gathering and storing information about site visitors, analyzing the information, and, based on the analysis, delivering the right information to each visitor at the right time. It is a key technology needed in various e-business applications. The elements of Personalization system includes

- Identify site visitor
- Retrieve visitor's profile(Id, Password, interest, role, business needs etc)
- Select content that matches visitor's preferences
- Retrieve content and assemble page for display to visitor

Principal elements of Web personalization include (a) the categorization and preprocessing of Web data, (b) the extraction of correlations between and across different kinds of such data, and (c) the determination of the actions that should be recommended by such a personalization system [13]. In this work we focus on Web usage mining. This process

relies on the application of statistical and data mining methods to the Web log data, resulting in a set of useful patterns that indicate users' navigational behavior. The data mining methods that are employed are: association rule mining, sequential pattern discovery, clustering, and classification. This knowledge is then used from the system in order to personalize the site according to each user's behavior and profile. Today, personalization is increasingly used as a means to expedite the delivery of information to a visitor, making the site useful and attractive to return to.

a) *Data Preprocessing*

An extensive description of data preparation and preprocessing methods can be found in. The data preprocess includes three basic steps like

- Data Cleaning
- User Identification
- Session Identification

The first issue in the preprocessing phase is data cleaning. Depending on the application, Web log data may need to be cleaned from entries involving pages that returned an error or graphics file accesses. In some cases such information might be useful, but in others such data should be eliminated from a log file.

Most important of all is the user identification issue. There are several ways to identify individual visitors. The most obvious solution is to assume that each IP address (or each IP address/client agent pair) identifies a single visitor. Nonetheless, this is not very accurate because, for example, a visitor may access the Web from different computers, or many users may use the same IP address (if a proxy is used). A further assumption can then be made, that consecutive accesses from the same host during a certain time interval come from the same user. More accurate approaches for a priori identification of unique visitors are the use of cookies or similar mechanisms or the requirement for user registration. However, a potential problem in using such methods might be the reluctance of users to share personal information.

The next step is to perform session identification, by dividing the click stream of each user into sessions. The usual solution in this case is to set a minimum timeout and assume that consecutive accesses within it belong to the same session, or set a maximum timeout, where two consecutive accesses that exceed it belong to different sessions.

b) *User Profiling*

User profiling is the process of collecting information about the characteristics, preferences, and activities of a Web site's visitors. This can be accomplished either explicitly or implicitly. Explicit collection of user profile data is performed through the use of online registration forms, questionnaires, and the like. The methods that are applied for implicit collection of user profile data vary from the use of cookies or similar technologies to the analysis of the users' navigational behavior that can be performed using Web usage mining techniques.

The extraction of information concerning the navigational behavior of Web site visitors is the objective of Web usage mining. Nevertheless this process can also be regarded as part of the creation of user profiles; it is therefore evident that those two modules overlap and are fundamental in the Web personalization process. A user profile can be either static, when the information it contains is never or rarely altered (e.g., demographic information), or dynamic when the user profile's data change frequently. Such information is obtained either explicitly, using online registration forms and questionnaires resulting in static user profiles, or implicitly, by recording the navigational behavior and/or the preferences of each user, resulting in dynamic user profiles. In the latter case, there are two further options: either regarding each user as a member of a group and creating aggregate user profiles, or addressing any changes to each user individually. When addressing the users as a group, the method used is the creation of aggregate user profiles based on rules and patterns extracted by applying Web usage mining techniques to Web server logs.

II. Related Works

Lamberti *et al.* proposed a relation-based page rank algorithm for Semantic Web search engines [10]. With the incredible increase of data available to end users through the Web, search engines come to play ever a more critical role. However, due to their general-purpose approach, it is always less uncommon that obtained result sets provide a burden of useless pages. The next-generation Web architecture [8], characterized by the Semantic Web, provides the layered architecture possibly allowing overcoming this limitation. Many search engines have been proposed, that allow increasing data retrieval accuracy by exploiting a key content of semantic Web resources, that is, relations. On the other hand, in order to rank results, the majority of the existing solutions need to work on the whole annotated knowledge base. In this paper, the author proposed a relation-based page rank algorithm to be used in conjunction with semantic Web search engines that simply relies on information that could be extracted from user queries and on annotated resources. Relevance is calculated as the probability that a retrieved resource actually contains those relations whose existence was assumed by the user at the time of query definition.

Ranking web pages using machine learning approaches is put forth by Sweah *et al* [19]. One of the key components which guarantee the acceptance of web search service is the web page ranker - a

component which is said to have been the main contributing factor to the early successes of Google. It is well recognized that a machine learning method such as the Graph Neural Network (GNN) can be able to learn and estimate Google's page ranking algorithm. This paper demonstrates that the GNN can successfully learn many other Web page ranking methods [7] e.g. TrustRank, HITS and OPIC. Experimental results illustrate that GNN may be suitable to learn any arbitrary web page ranking scheme, and hence, may be more flexible than any other existing web page ranking scheme. The significance of this inspection lies in the fact that it is possible to learn ranking schemes for which no algorithmic solution exists or is known.

Shohel Ahmed *et al.* proposed a personalized URL re-ranking method based on psychological characteristics of users browsing like “common-mind,” “uncommon-mind,” and “extremely uncommon-mind” [17]. This personalization method constructs an index of the anchor text retrieved from the web pages that the user has clicked during his/her past searches. Our method provides different weights to the anchor text according to the psychological characteristics for re-ranking URLs.

Srour *et al.* provided a personalized Web Page ranking using trust and similarity. Search engines, like Google, utilize link structure to rank web pages [18]. Although this technique offers an objective global estimate of the web page importance, it is not targeted to the specific user preferences. This paper presents a new technique for the personalization of the results of a search engine based on the user's taste and preferences. The idea of trust and similarity, obtained from explicit user input and implicit user behavioral patterns, are used to compute personalized page rankings [6].

Shiguang *et al.* given the improvement of page ranking algorithm based on timestamp and link [16]. The conventional ranking technique favors the old pages, which makes old pages always emerge on the top of the ranking results when pages are ranked according to the dynamic web by the static ranking algorithm. Therefore, this paper proposes a temporal link - analysis technique to overcome the problem. This technique uses the last variation time that returned by the HTTP response as the timestamp of nodes and links concerned. And the weight of the in-link and out-link are also combined to calculate the overall weight of the pages. Using the WTPR technique can make the old pages decline and new pages rise in the ranking result, meanwhile it can help the old pages which have high-quality get higher rank value than common old pages.

Kritikopoulos *et al.* proposed Wordrank: a method for ranking web pages based on content similarity [9]. This paper presents WordRank, a new page ranking system, which utilize similarity between interconnected pages. WordRank establishes the model of the biased surfer which is based on the following hypothesis: "the visitor of a Web page have a tendency to visit Web pages with similar content rather than content irrelevant pages". This technique modifies the random surfer model by biasing the probability of a user to follow a link in favor of links to pages with similar content. It is the perception that WordRank is most suitable in topic based searches, since it prioritizes strongly interconnected pages, and in the same time is more robust to the multitude of topics and to the noise produced by navigation links. This paper provides preliminary experimental verification from a search engine developed for the Greek fragment of the World Wide Web.

Magdalini Eirinaki *et al.* present a hybrid probabilistic predictive model extending the properties of Markov models by incorporating link analysis methods [12]. More specifically, we propose the use of a PageRank-style algorithm for assigning prior probabilities to the web pages based on their importance in the web site's graph. We prove, through experimentation, that this approach results in more objective and representative predictions than the ones produced from the pure usage-based approaches.

III. Markov Models

The 1st-order Markov models (Markov Chains) provide a simple way to capture sequential dependence [3], but do not take into consideration the “long-term memory” aspects of web surfing behavior since they are based on the assumption that the next state to be visited is only a function of the current one. Higher-order Markov models [11] are more accurate for predicting navigational paths, there exists, however, a trade-off between improved coverage and exponential increase in statespace complexity as the order increases. Moreover, such complex models often require inordinate amounts of training data, and the increase in the number of states may even have worse prediction accuracy and can significantly limit their applicability for applications requiring fast predictions, such as web personalization. There have also been proposed some mixture models that combine Markov models of different orders. Such models, however, require much more resources in terms of preprocessing and training. It is therefore evident that the final choice that should be made concerning the kind of model that is to be used, depends on the trade-off between the required prediction accuracy and model’s complexity/size.

Hidden markov model (HMM) are generative, directed graphical models, which describe the joint probability over a state sequence and output sequence. Such generative models make limiting independence assumptions over the output sequence.

Mixed Markov models are based on the selection of parts from Markov models of different order,

so that the resulting model has reduced state complexity as well as increased precision in predicting the user's next step. Deshpande and Karypis *et al.* [5] propose the All-Kth-Markov models, presenting 3 schemas for pruning the states of the All-K*th*-Order Markov model. Cadez *et.al* [4] as well as Sen and Hansen *et al.* [15] also proposed the use of mixed Markov models. A different approach is that of [1] Acharyya and Ghosh *et al*, who use concepts, to describe the web site. Each visited page is mapped to a concept, imposing a tree hierarchy on these topics. A semi-Markov process is then defined on this tree based on the observed transitions among underlying visited pages. They prove that this approach is computationally much less demanding compared to using higher order Markov models.

IV. Conclusion

The explosive growth of information sources available on the World Wide Web has necessitated the users to make use of automated tools to locate desired information resources and to follow and asses their usage pattern. Web page prefetching has been widely used to reduce the access latency problem of the internet, its success mainly relies on the accuracy of web page prediction. Markov model is the most commonly used prediction model because of its high accuracy. Low order markov models have higher accuracy and lower coverage. The higher order models have a number of limitations associated with

i) Higher state complexity,
ii) Reduced coverage,
iii) Sometimes even worse prediction accuracy.

To overcome these limitations of higher order markov model by Hidden markov model. It is a powerful method for labeling sequence data but it has two major drawbacks such as one stemming from its independence assumptions and the other from its generative nature.

We have discussed some of the techniques to overcome the issues of web page change ranking. As the web is going to expand, web usage in web databases will become more and more and the rank prediction is also more difficult. The above findings will become will be good guide to rank the web pages effectively. In this paper, we have presented a comprehensive survey of up-to-date researchers of ranking web pages for web personalization. Besides, a brief introduction about web mining, web personalization and web page change ranking have also been presented. However, research of the web page ranking is just at its beginning and much deeper understanding needs to be gained.

V. Future Work

This survey paper intends to aid upcoming researchers in field of web page ranking for web personalization to understand the available methods and help to perform their research in right direction. For future work, there are some improvements that can be implemented. First, the first-order Markov models (Markov Chains) provide a simple way to capture sequential dependence, but do not take into consideration the "long-term memory" aspects of web surfing behavior since they are based on the assumption that the next state to be visited is only a function of the current one. Higher-order Markov models and hidden markov models are more accurate for predicting navigational paths, there exists, however, a trade-off between improved coverage and exponential increase in state space complexity as the order increases. Secondly, to predict web page ranks efficiently by doing the preprocessing phase effectively.

References Références Referencias

1. S. Acharyya and J. Ghosh. "Context-Sensitive Modeling of Web- Surfing Behaviour Using Concept Trees, *in Proc. Of the 5th WEBKDD Workshop,* Washington DC, August 2003.
2. M. S. Aktas, M.A. Nacar and F. Menczer. Personalizing Page Rank Based on domain Profiles, *Processing of WEBKDD 2004* Workshop, 2004.
3. S. Brin and L.Page. The anatomy of a large-scale hypertextual Web search engine, *in Proc. of the 7th International World Wide Web Conference (WWW7),* Brisbane, 1998.
4. I. Cadez, S. Gaffney and P.Smyth. A general probabilistic framework for clustering individuals and objects*, in Proc. Of the 6th ACM SIGKDD Conference*, Boston, 2000.
5. M. Deshpande, and Karypis. Selective Markov Models for Predicting Web-Page Accesses, *Proc. of the 1st SIAM International Conference on Data Mining*, 2001.
6. T. H Haveliwala. Topic-sensitive PageRank, *Processing of WWW,* 2002.
7. H. Y kao and S.Flin. A Fast PageRank Convergence Method based on the Cluster rediction, *Proc. Of IEEE/WIC/ACM International Conference on Web Intelligence*, 2007.
8. M. Y Kan and H.O.N Thi. Fast webpage classification using URL features, Processing of CIKM, 2005.
9. A.Kritikopoulos, M. Sideri and I. Varlamis. WordRank: A Method for Ranking Web Pages Based on Content Similarity, *British National Conference on Databases,* pp. 92-100, 2007.

10. F.Lamberti, A. Sanna, and C. Demartini. A Relation-Based Page Rank Algorithm for Semantic Web Search Engines, *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 123-136, 2009.
11. M. Levene and G. Loizou. Computing the Entropy of User Navigation in the Web, *in Intl. Journal of Information Technology and Decision Making,* 2:459-476, 2003.
12. Magdalini Eirinaki, Michalis Vazirgiannis, Dimitris Kapogiannis. Web path Recommendations Based on Page Ranking and Markov Models, *WIDM'05*, November 05, 2005.
13. B. Mobasher, R. Cooley and J.Srivastava. Automatic personalization based on web usage mining,. *Commun. ACM,* 43, 8 (August), 142–151, 2000a.
14. M. D. Mulvenna, S. S. Anand and A.G Buchner. Personalization on the net using web mining,. *Commun*. ACM, 43, 8 (August), 123–125, 2000.
15. R. Sen and M. Hansen. Predicting a Web user's next accessbased on log data, *in Journal of Computational Graphics and Statistics,* 12(1):143-155, 2003.
16. Shiguang Ju, Zheng Wang and Xia Lv. Improvement of Page Ranking Algorithm Based on Timestamp and Lin, *International Symposiums on Information Processing* (ISIP), pp. 36-40, 2008.
17. Shohel Ahmed, Sungjoon Park, Janson, J. Jung and Sanggil Kang. A Personalized URL Re-ranking ethod using Psychological User Browsing Characteritics, *Journal of Universal Computer Science*, vol.15, no.4,2009.
18. L. Srour, A. Kayssi and A. Chehab. Personalized Web Page Ranking Using Trust and Similarity, *in Intl. Conference on Computer Systems and Applications*, pp. 454-457, 2007.
19. Sweah Liang Yong, M. Hagenbuchner. M and ah chung Tsoi. Ranking Web Pages Using Machine Learning Approaches*, in Intl. Conference on Web Intelligence and Intelligent Agent Technology,* vol. 3, pp. 677-680, 2008.

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

A Study of Spam E-mail classification using Feature Selection package

By R. Parimala, Dr. R. Nallaswamy
National Institute of Technology

Abstract- Feature selection (FS) is the technique of selecting a subset of relevant features for building learning models. FS algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset. Many FS algorithm have been proposed. This paper presents a new FS technique which is guided by Fselector Package. The package Fselector implements a novel FS algorithm which is devoted to the feature ranking and feature subset selection of high dimensional data. This package provides functions for selecting attributes from a given dataset. Attribute subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. The R package provides a convenient interface to the algorithm. This paper investigates the effectiveness of twelve commonly used FS methods on spam data set. One of the basic popular methods involves filter which select the subset of feature as preprocessing step independent of chosen classifier, Support vector machine classifier. The algorithm is designed as a wrapper around five classification algorithms. The short description of the algorithm and performance measure of its classification is presented with the spam data set.

Keywords: *FS, filter, wrapper, best-first search, SVM classification.*

GJCST Classification: *H.2.8, F.2.2*

Strictly as per the compliance and regulations of:

A Study of Spam E-mail classification using Feature Selection package

R. Parimala[α], Dr. R. Nallaswamy[Ω]

Abstract- Feature selection (FS) is the technique of selecting a subset of relevant features for building learning models. FS algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset. Many FS algorithm have been proposed. This paper presents a new FS technique which is guided by Fselector Package. The package Fselector implements a novel FS algorithm which is devoted to the feature ranking and feature subset selection of high dimensional data. This package provides functions for selecting attributes from a given dataset. Attribute subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. The R package provides a convenient interface to the algorithm. This paper investigates the effectiveness of twelve commonly used FS methods on spam data set. One of the basic popular methods involves filter which select the subset of feature as preprocessing step independent of chosen classifier, Support vector machine classifier. The algorithm is designed as a wrapper around five classification algorithms. The short description of the algorithm and performance measure of its classification is presented with the spam data set.

Keywords-FS, filter, wrapper, best-first search, SVM classification.

I. Introduction

Classification is a method of categorizing or assi--gning class labels to a pattern set under the sup--ervision of teacher. It is one of the familiar and popular techniques in machine learning. The decisions boundaries are generated to discriminate between patterns belong to different classes. The patterns are initially partitioned into training set and testing set randomly and the classifier is trained on the former. The testing set is used to evaluate the generalized capability of the classifier. When a classification problem has to be solved, the common approach is to compute a wide variety of features that will carry as much as possible different information to perform the classification of samples. Thus, numerous features are used whereas, generally, only a few of them are relevant for the classification task, including the other in the feature set used to represent the samples to classify, may lead to a slower execution of the classifier, less understandable results, and much reduced accuracy[1]. The irrelevant features are filtered out before the classification process[1]. Their main advantage is that their low computational complexity which makes them very fast. Their main drawback is that they are not optimized to be used with a particular classifier as they are completely independent of the classification stage.

II. Related Work

Kira and Rendell (1992) described a statistical feature selection algorithm called RELIEF that uses instance based learning to assign a relevance weight to each feature [2][3]. John, Kohavi and Pfleger (1994) addressed the problem of irrelevant features and the subset selection problem. Further, they claim that the filter model approach to subset selection should be replaced with the wrapper model [4]. Koller and Sahami (1996) examined a method for feature subset selection based on Information Theory: they presented a theoretically justified model for optimal feature selection based on using cross-entropy to minimize the amount of predictive information lost during feature elimination [5]. Dash and Liu (1997) gave a survey of feature selection methods for classification. In a comparative study of feature selection methods in statistical learning of text categorization (with a focus is on aggressive dimensionality reduction)[34], Yang and Pedersen (1997) evaluated document frequency (DF), information gain (IG), mutual information (MI), a χ^2 test (CHI) and term strength (TS); and found IG and CHI to be the most effective[20]. Kohavi and John (1997) introduced wrappers for feature subset selection[4]. Their approach searches for an optimal feature subset tailored to a particular learning algorithm and a particular training set. Xing, Jordan and Karp (2001) successfully applied feature selection methods (using a hybrid of filter and wrapper approaches) to a classification problem.

Naïve Bayes Network algorithms were used frequently and they have shown a considerable success in filtering English spam e-mails [1]. Knowledge-based and rule-based systems were also used by researchers for English spam filters [2] [3]. As an alternative to these classical learning paradigms used frequently in spam filtering domain, evolutionary method was employed for classification and compared with Naïve Bayes

About[α]- R. Parimala, Assistant professor in Computer science Department, Periyar E.V.R. College, Tiruchirapalli, India. (email: parimadhu2003@yahoo.com).

About[Ω] - Dr. R. Nallaswamy, Profeesor, Department of Mathematics, National Institute of Technology, Tiruchirapalli, India. (email:nalla@nitt.edu).

classification [4]. It was argued that they show similar success rates although the former outperforms the Naïve Bayes classifier in terms of speed.

III. Backgrounds

In this section, we discuss the basic concepts related to our research. Topics include a brief background on FS, methods, Feature Ranking and Feature subset Algorithms.

IV. Feature Selection

FS is frequently used as a preprocessing step to machine learning. It is a process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criterion. FS has been a fertile field of research and development since 1970's and proven to be effective in removing irrelevant and redundant features, increasing efficiency in learning tasks, improving learning performance like predictive accuracy, and enhancing comprehensibility of learned results[4]. In recent years, data has become increasingly larger in both the number of instances and the number of features in many applications.

V. Feature Selection Methods

Techniques for FS can be divided in two approaches: **feature ranking** and **subset selection**. In the first approach, features are ranked by some criteria and then features above a defined threshold are selected. In the second approach, one searches a space of feature subsets for the optimal subset. Moreover, FS methods can broadly fall into two broad categories, the filter model or the wrapper model [2]. The filter model relies on general characteristics of the training data to select some features without involving any learning algorithm. The wrapper model requires one pre determined learning algorithm in FS and uses its performance to evaluate and determine which features are selected. As for each new subset of features, the wrapper model needs to learn a hypothesis (or a classifier). It tends to find features better suited to the predetermined learning algorithm resulting in a superior learning performance, but it also tends to be more computationally expensive than the filter model [5]. When the number of features becomes very large the filter model is usually chosen due to its computational efficiency.

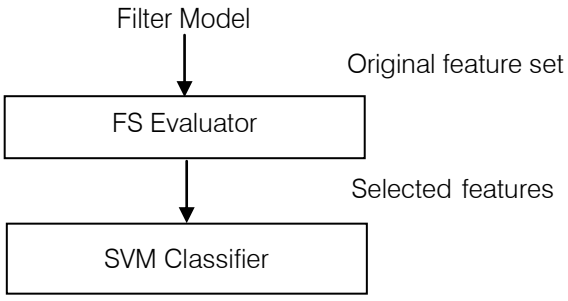


Fig 1. Filter Model

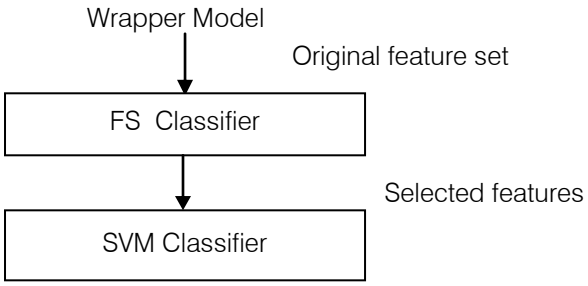


Fig 2. Wrapper Model

In wrapper approaches learning algorithms are used to evaluate the quality of each feature. Specifically, a learning algorithm is run on a feature subset, and the classification accuracy of the feature subset is taken as a measure for feature quality. Generally, wrapper approaches are more computational demanding as compared with filter approaches. However, wrapper approaches often are superior in accuracy when compared with filters approaches which ignore the properties of the learning task in hand. In most application of SVM classification tasks, accuracy plays a greater role as compared with that of computational cost. Both approaches, filters and wrappers, usually involve combinatorial searches through the space of possible feature subsets. In the past few decades, researchers have developed large amount of FS algorithms. These algorithms are designed to serve different purposes, are of different models, and all have their own advantages and disadvantages. Various feature ranking and FS techniques have been proposed such as Correlation-based FS (CFS), Chi-square Feature Evaluation, Information Gain (IG), Gain Ratio (GR), Symmetric Uncertainty (SU), oneR and ReliefF. The feature ranking algorithms are implemented based on the code from Fselector package. The FSelector Package was created by Piotr Romanski and released in April 11, 2009.

VI. Feature Ranking Approach

The primary purpose of feature ranking approach is to reduce the dimensionality to decrease the computation time. This is particularly important concerning text categorization where the high dimensionality of the feature space is a problem. In many cases the number of features is in the tens of thousands. Then it is highly desirable to reduce this

number, preferably without any loss in accuracy. Several FS methods have been proposed.
The general algorithm for the Feature Ranking Approach is:

```
for each feature Fi
  wfi = getFeatureWeight(Fi)
  add wfi to wt_list
sort wt_list
choose top-k features.
```

a) *Correlation based FS (CFS)*

CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Yang & Pedersen, 1997 is used to measure the association between a class and features, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation[1]. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best-first search and genetic search. Among given features, it finds out an optimal subset which is best relevant to a class having no redundant feature. It evaluates merit of the feature subset on the basis of hypothesis--"Good feature subsets contain features highly correlated with the class, yet uncorrelated to each other [7]". This hypothesis gives rise to two definitions. One is feature class correlation and another is feature-feature correlation. Feature-class correlation indicates how much a feature is correlated to a specific class while feature-feature correlation is the correlation between two features. Equation 1, also known as Pearson's correlation, gives the merit of a feature subset consisting of k number of features. The CFS method is based on the "merit" criterion.
Equation for CFS is given is equation

$$r_{zc} = \frac{k\,\bar{r}_{zi}}{\sqrt{k + k(k-1)\bar{r}_{ii}}} \qquad (1)$$

where r_{zc} is the correlation between the summed feature subsets and the class variable, k is the number of subset features, $\bar{r}_{zi}$ is the average of the correlations between the subset features an the class variable, and $\bar{r}_{ii}$ is the average inter-correlation between subset features[7]. In CFS features can be classified into three disjoint categories, namely, strongly relevant, weakly relevant and irrelevant features [4]. Strong relevance of a feature indicates that the feature is always necessary for an optimal subset; it cannot be removed without affecting the original conditional class distribution. Weak relevance suggests that the feature is not always necessary but may become necessary for an optimal subset at certain conditions. Irrelevance indicates that the feature is not necessary at all.

b) *CHI (χ 2 statistic)*

Chi-Squared attribute selection is based on the Chi-Squared Statistic with respect to the target class. The algorithm finds weights of discrete attributes basing on a chi-squared test. The *χ2 test is used in statistics to test the independence* between two events [6].

c) *EN (Entropy-based Ranking)*

Linear correlation may not be able to capture correlations that are not linear. Therefore non-linear correlation measures often adopted for measurement. It is based on the information-theoretical concept of entropy, a measure of the uncertainty of a random variable.

d) *IG (Information Gain)*

Information gain [27], of a term measures the number of bits of information obtained for category prediction by the presence or absence of the term in a document. Information Gain is a method that selects attributes based on informational value gained by creating a branch on the attribute with respect to the class. Information theory indices are most frequently used for feature evaluation. A probabilistic model of a nominal valued feature Y can be formed by estimating the individual probabilities of the values yεY from the trained data. Entropy is a measure of uncertainty or unpredictability in a system. The entropy of Y is given by $H(Y) = -\sum_{y \in Y} P(y)\log_2(p(y))$. If the observed value of Y in the training data are partitioned according to the value of a second feature x, and the entropy of Y with respect to the partitions induced by x is less than the entropy of Y prior to partitioning, then there is a relationship between feature Y and x. The entropy of Y after observing x is

$$H(Y/x) = -\sum_{x \in X} p(x) \sum_{y \in Y} P(y/x)\log_2(p(y))$$

Information gain is given by

$$Gain = H(Y) - H(Y/x)$$

$$= \mathrm{H(X)} - \mathrm{H(X/Y)}$$

$$= H(y) + H(x) - H(x,Y)$$

Information gain is a symmetrical measure. The amount of information gained about y after observing x is equal to the amount of information gained about x after observing y.

e) *Gain Ratio*

Gain Ratio is a modification to information gain that takes into account the number and size of daughter nodes into which an attribute splits the dataset with respect to the class. This dampens the preference that the information gain method has for attributes with large numbers of possible values. [8]

$$Gain\ \ Ratio = \frac{H(Y)+H(X)-H(Y,X)}{H(X)}.$$

f) *Mutual Information*

The MIFS (Mutual Information FS) algorithm uses a forward selection (Battiti, 1994). Mutual Information is a measure of general interdependence between random variables (i.e., features and type).We define the mutual information, I[X; Y] ,

I[X; Y] = H[X] - H[X/Y]

= H[Y] = H[Y /X]

= H[Y] + H[X] - H[X; Y]

g) *Symmetrical Uncertainty*

Symmetrical Uncertainty is another method that was devised to compensate for information gain's bias towards features with more values. It capitalizes on the symmetrical property of information gain. The symmetrical uncertainty between features and the target concept can be used to evaluate the goodness of features for classification [10]

$$\text{Symmetrical uncertainty} = 2\frac{Gain}{H(Y)+H(X)}$$

h) *OneR*

OneR could be viewed as an extremely powerful filter, reducing all datasets to one feature. OneR algorithms find weights of discrete attributes basing on very simple association rules involving only one attribute in condition part. The algorithm uses OneR classifier to find out the attributes' weights. For each attribute it creates a simple rule based only on that attribute and then calculates its error rate [11].

i) *Relief*

The RELIEF, one of the most used filter methods was introduced by Kira and Rendell [2] In the RELIEF, the relevance weight of each feature is estimated according to its ability to distinguish instances belonging to different classes. Thus, a good feature must assume similar values for instances in the same class and different values for instances in other classes. The algorithm finds weights of continuous and discrete attributes basing on a distance between instances. The relevance weights are set to be zero for each feature and then are estimated iteratively. In order to do that, an instance is chosen randomly from the training dataset. Then, the RELIEF searches for two closest neighbors to such instance, one in the same class, called the *Nearest Hit* and the other in the opposite class called the *Nearest Miss*. The relevance weight of each feature is modified in each step according to the distance of the instance to its *Nearest Hit* and *Nearest Miss*. The relevance weights continue to be updated by repeating the above process using a random sample of *n* instances drawn from the training dataset. Filter methods are fast but lack of robustness against interactions among features and feature redundancy. In addition, it is not clear how to determine the cut-off point for rankings to select only truly important features and exclude noise. ReliefF uses a nearest neighbor implementation to maintain relevancy scores for each attribute. It defines a good discriminating attribute as the attribute that has the same value for other attributes in the same class and different from attribute values in different classes. [7][8][9] The Weka implementation repeatedly evaluates an attribute's worth by considering the value of its n nearest neighbors of same and different classes. [4] A family of algorithms called Relief [4] is based on the feature weighting, estimating how well the value of a given feature helps to distinguish between instances that are near to each other. One advantage of Relief is that it is sensitive to feature interactions and can detect higher than pair wise interactions.

VII. Feature Subset Selection Approach

Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model. Wrapper methods search through the space of feature subsets and calculate the estimated accuracy of a single learning algorithm for each feature that can be added to or removed from the feature subset. The feature space can be searched with various strategies, e. g., forwards (i. e., by adding attributes to an initially empty set of attributes) or backwards (i. e., by starting with the full set and deleting attributes one at a time). Usually an exhaustive search is too expensive, and thus non-exhaustive, heuristic search techniques like genetic algorithms, greedy stepwise, best first or random search are often used (see, for details, Kohavi and John (1997)). For extracting the wrapper subsets we used wrapper subset evaluator in combination with the best first search method. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated.

In the feature subset selection approach, one searches a space of feature subsets for the optimal subset. Such approach is present on the FSelector package by wrappers techniques (e.g. best-first search,

backward search, forward search, hill climbing search). Those techniques works by informing a function that takes a subset and generate an evaluation value for that subset. A search is performed in the subsets space until the best solution can be found.

a) *Feature Subset Selection Algorithm*

The feature subset algorithm conducts a search for a good subset using the induction algorithm itself as part of the evaluation function. The accuracy of the induced classifiers is estimated using accuracy estimation techniques [4]. The wrapper approach conducts a search in the space of possible parameters. Wrapper approaches use a specific machine learning algorithm/classifiers and utilize the corresponding classification performance to select features. A search requires a state space, an initial state, a termination condition, and a search engine [15]. Best-first search is a more robust method than hill-climbing. The idea is to select the most promising node we have generated so far that has not already been expanded. Best-first search usually terminates upon reaching the goal.

b) *Searching the Feature Subset Space*

The purpose of FS is to decide which of the initial (possibly large number) of features to include in the final subset and which to ignore. If there are n possible features initially, then there are 2n possible subsets. The only way to find the best subset would be to try them all---this is clearly prohibitive for all but a small number of initial features.

Various heuristic search strategies such as hill climbing and Best First [Rich and Knight, 1991] are often applied to search the feature subset space in reasonable time. Two forms of hill climbing search and a Best First search were trialed with the feature selector described below; the Best First search was used in the final experiments as it gave better results in some cases. The Best First search starts with an empty set of features and generates all possible single feature expansions. The subset with the highest evaluation is chosen and is expanded in the same manner by adding single features. If expanding a subset results in no improvement, the search drops back to the next best unexpanded subset and continues from there. Given enough time a Best First search will explore the entire search space, so it is common to limit the number of subsets expanded that result in no improvement. The best subset found is returned when the search terminates[12]. The general algorithm for the Feature Subset Selection approach is:

S = all subsets
for each subset s in S
  evaluate(s)
return (the best subset).

1) *LDA*

Linear discriminant analysis (LDA) and the related Fisher's linear discriminant are methods used in statistics and machine learning to find a linear combination of features which characterize or separate two or more classes of objects or events. The resulting combination may be used as a linear classifier or, more commonly, for dimensionality reduction before later classification. The LDA problem is formulated as follows . Let $x \in \Phi^n$ be a feature vector. We seek to find a transformation $\bar{x} = \theta x$, $\theta : \Phi^n \to \Phi^m$ with $m < n$, such that in the transformed space, minimum loss of discrimination occurs. In practice, m is much smaller than n. A common form of optimality criteria to be maximized is the function $J = tr(S_W^{-1} S_B)$. In classical LDA, the corresponding input-space within-class and between-class scatter matrix are defined by,

$$S_B = \sum_{k=1}^{K} n_k (v_k - v)(v_k - v)^t$$

$$S_W = \sum_{k=1}^{K} \sum_{n=1}^{n_k} (x_n^k - v_k)(x_n^k - v_k)^t$$

$$v_k = \frac{1}{n_k} \sum_{n=1}^{n_k} x_n^k$$

$$v = \frac{1}{N} \sum_{k=1}^{K} n_k v_k$$

The LDA is to maximize in some sense the ratio of between-class and within-class scatter matrices after transformation. This will enable to choose a transform that keeps the most discriminative information while reducing the dimension. Precisely, we want to maximize the objective function

$$\max_{\theta} \frac{\left| \theta S_B \theta^t \right|}{\left| \theta S_w \theta^t \right|}$$

The columns of the optimum θ are the relative generalized eigenvectors corresponding to the first p maximal magnitude eigenvalues of the equation $S_B \mu = \lambda S_w \mu$ [13].

2) *Random Forest*

Random forest (or RF) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. Random forests are often used when we have very large training datasets and a very large number of input variables (hundreds or even thousands of input variables). A random forest model is typically made up of tens or hundreds of decision trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler [14].

3) *RPART*

Recursive PARTitioning is a fundamental tool in data mining. Classification and regression trees [18] can be generated through the **rpart** package [19]. The rpart programs build classification or regression models of a very general structure using a two stage procedure; the resulting models can be represented as binary trees. The tree is built by the following process: first the single variable is found which best splits the data into two groups The data is separated, and then this process is applied separately to each sub-group and so recursively until the subgroups either reach a minimum size or until no improvement can be made.

4) *NAÏVE BAYES*

The Naïve Bayes (NB) classifier is the simplest in terms of its ease of implementation [20]. In terms of a classifier Bayes theorem (4) can be expressed as

$$P(C/F)=\frac{P(F/C)P(C)}{P(F)}$$,where F is a set of features and C are the target class. One argument [35] is that with the independence assumption the classifier would produce poor probabilities, but the ratio between them would be approximately the same as using conditional probabilities. Using the somewhat 'Naive' independence assumption gave birth to its name Naive Bayesian classifier. Using the assumption for independence, according to (1), the joint probability for all n features can be obtained as a product of the total individual probabilities.

$$P(F/C)=\prod_{i=1}^{n}P(f_i/C)$$

$$P(C/F)=\frac{P(C)\prod_{i=1}^{n}P(f_i/C)}{P(F)}$$

The denominator P(F) is the probability of observing the features in any message and can be expressed as

$$P(F)=\sum_{k=1}^{m}P(C_k)\prod_{i=1}^{n}P(f_i/C_k)$$

Inserting (8) into (7) the formula used by the Naive Bayesian Classifier is obtained

$$P(C/F)=\frac{P(C)\prod_{i=1}^{n}P(f_i/C)}{\sum_{k=1}^{m}P(C_K)\prod_{i=1}^{n}P(f_i/C_k)}$$

5) *SVM*

SVM [18][19] separates two classes with vectors that pass through training data points. The separation is measured as the distance between the support vectors and is called the margin. SVM have shown promising results concerning text categorization problems in several studies [20]. A recent study [21] demonstrated that its performance was good with reference to the spam domain.

Support vector machine and its parameters

The algorithm about SVM is originally established by Vapnik (1998). Since 1990s SVM has been a promising tool for data classification. This introduction to Support Vector Machines (SVMs) is based on [26], [27], [28] and [29]. Support vector machine [22], [23] has gained prominence in the field of machine learning. Its basic idea is to map data into a high dimensional space and find a separating hyper plane with the maximal margin[22][23]. The solutions to classification sought by kernel based algorithm such as the SVM are linear functions in feature space: $\mathrm{f(x)}=\mathrm{w}^{\mathrm{T}}\phi(x)$ for some weight vector $\mathrm{w}\in\mathrm{F}$.

Given a training set of instance-label pairs (x_i,y_i), i =1, 2, 3…ℓ, where $x_i\in R^n$. The class label of the i[th] pattern is denoted by $y_i\in\{1,-1\}^t$. Nonlinearly separable problem are often solved by mapping the input data samples x_i to a higher dimensional feature space $\phi(x_i)$. The classical maximum margin SVM classifier aims to find a hyper plane of the form $w^t\phi(x)+b=0$ that separates patterns of the two classes[30]. So far we have restricted ourselves to the case where the two classes are noise-free. In the case of noisy data, forcing zero training error will lead to poor generalization. To take account of the fact that some data points may be misclassified we introduce a vector of slack variables $\Xi=(\xi_1,\ldots,\xi_l)^T$ that measure the amount of violation of the constraints. The problem can then be written as

$$\underset{w,b,\xi}{Minimize}\frac{1}{2}w^tw+C\sum_{i=1}^{n}\xi_i \qquad (2)$$

Subject to the constraints

$$y_i\left(w^t\phi(x_i)+b\right)\geq 1-\xi_i$$

$$\xi_i\geq 0, i=1,2,3.....\ell, \qquad (3)$$

The solution to (2)-(3) yields the soft margin classifier, so termed because the distance or margin between the separating hyper plane $w^t\left(\phi(x)+b\right)=0$ is usually determined by considering the dual problem, which is given by

$$L(w,b,a_i,\mathrm{E},\Gamma)=\frac{\|\mathrm{w}\|^2}{2}+\sum_{\mathrm{i}=1}^{\ell}\alpha_{\mathrm{i}}\left[\mathrm{y}_{\mathrm{i}}\left(\mathrm{w}^{\mathrm{T}}\varphi(\mathrm{x}_{\mathrm{i}})+\mathrm{b}_{\mathrm{i}}\right)-1+\xi_{\mathrm{i}}\right]-\sum_{\mathrm{i}=1}^{\ell}\gamma_{\mathrm{i}} \qquad (4)$$

where $\Lambda=(\alpha_1,\ldots,\alpha_l)^T$, as before, and $\Gamma=(\gamma_1,\ldots,\gamma_l)^T$ are the Lagrange multipliers

corresponding to the positivity of the slack variables. The solution of this problem is the saddle point of the Lagrangian given by minimizing L with respect to $\mathbf{w}, \Xi$ and b, and maximizing with respect to $\Lambda \geq 0$ and $\Gamma \geq 0$. Differentiating with respect to $\mathbf{w}$, b and Ξ and setting the results equal to zero.

We obtain

$$\frac{\partial L(\mathbf{w}, b, \alpha, \Xi, \Gamma)}{\partial \mathbf{w}} \quad \mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i) = 0$$

$$\frac{\partial L(\mathbf{w}, b, \alpha, \Xi, \Gamma)}{\partial b} \quad -\sum_{i=1}^{l} \alpha_i y_i = 0,$$

d

$$\frac{\partial L(\mathbf{w}, b, \Lambda, \Xi, \Gamma)}{\partial \xi_i} = C - \alpha_i - \gamma_i = 0. \qquad (4)$$

$$\underset{\alpha}{Minimize} \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y_i y_j \alpha_i \alpha_j k(x_i, x_j) - \sum_{i=1}^{\ell} \alpha_i$$

to

$$\sum_{I=1}^{\ell} \alpha_I d_i = 0$$

and $0 \leq \alpha_i \leq C, i = 1,2,3.....\ell$

Here, $\alpha_i \quad i = 1,2,3,.....\ell$ denotes the Lagrange multipliers and the matrix $K(x_i, x_j) = \phi(x_i).\phi(x_j)$ are termed as Kernel matrix. Kernel based learning methods use an implicit mapping of the input data into a high dimensional feature space defined by a kernel function. Training vector x_i is mapped into a higher dimensional feature space and then the learning takes place in the feature space [24][25]. In this paper, we focus our attention to the RBF kernels: $K(x_i, x_j) = \phi(x_i).\phi(x_j)$.

Package kernlab [27, 28] aims to provide the R user with basic kernel functionality (e.g., like computing a kernel matrix using a particular kernel), along with some utility functions commonly used in kernel-based methods like a quadratic programming solver, and modern kernel-based algorithms based on the functionality that the package provides.

ksvm() in kernlab package [27, 28] is a flexible SVM implementation which includes the most SVM formulations and kernels and allows for user defined kernels as well. It provides many useful options and features like a method for plotting, class probabilities output, cross validation error estimation.

VIII. Experimental Results

a) *K-Fold Cross Validation*

When we have finished the FS, we use the SVM to do the classification. The cross validation will help to identify good parameters so that the classifier can accurately predict unknown data. In this paper, we used 10 fold cross validation to choose the penalty parameter C and γ in the SVM. When we get the nice arguments, we will use them to train model and do the final prediction [33].

b) *Used Environment and Libraries*

There are several libraries available for FS and SVMs. Fselector package provides functions for selecting attributes from a given dataset. Attribute subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. This package contains Algorithms for filtering attributes, Algorithms for wrapping classifiers and search attribute subset space such as best first search, backward search, forward search and hill climbing search and Algorithm for choosing a subset of attributes based on attributes' weights.

The environment used in this work is R [30] together with the package kernlab [27][28]. Kernlab is a package that offers several methods for kernel-based learning. The program was written in R programming language. The PC we used for experiment has the machine used was an Intel Core 2 Duo E7500 @ 2.93GHz with 2GB RAM.

c) *Datasets and Data Preprocessing*

The data of the spam email problem in this paper is downloaded from the UCI Machine Learning Repository [31][32]. There are a total of 4601 emails in the database, i.e., the training set is of size 4601, 1813 of which are labeled as spam, the rest as non-spam. In addition to this class label there are 57 variables indicating the frequency of certain words and characters in the e-mail. The first 48 variables contain the frequency of the variable name (e.g., business) in the e-mail. If the variable name starts with num (e.g., num650) it indicates the frequency of the corresponding number (e.g., 650). These words were deemed to be relevant for distinguishing between spam and non-spam emails. They are as follows: make, address, all, 3d, our, over, remove, internet, order, mail, receive, will, people, report, addresses, free, business, email, you, credit, your, font, 000, money, hp, hpl, george, 650, lab, labs, 857, data, 415, 85, technology, 1999, parts, pm, direct, cs, meeting, original, project, re, edu, table, and conference. The variables 49-54 indicate the frequency of the characters ';', '(', '[', '!', '$', and '#'. The variables 55-57 contain the average; longest and total run-length of capital letters. Variable 58 indicates the type of the mail and is either "non-spam" or "spam", i.e. unsolicited commercial e-mail. . Given an email text and a particular WORD, we calculate its frequency, i.e., the percentage of words in the e-mail that match WORD: word freq WORD =100 × r/t, where r is number of times the

WORD appears in the email and t is the total number of words in e-mail.

In order to obtain an averaged unbiased accuracy estimate, we conducted 25 runs. For each run, data are completely randomized, then the database is divided into a training set and a separate test set.

d) *Measuring the performance*

The meaning of a good classifier can vary depending on the domain in which it is used. For example, in spam classification it is very important not to classify legitimate messages as spam as it can lead to e.g. economic or emotional suffering for the user. Classifiers have long been evaluated on their accuracy only. An often-used measure in the information retrieval and natural language processing communities is Overall Accuracy (OA). This is the most common and simplest measure to evaluate a classifier. It is just defined as the degree of right predictions of a model. Kappa statistic: (Kappa). This is originally a measure of agreement between two classifiers (Cohen, 1960), although it can also be employed as a classifier performance measure. This is the overall Accuracy corrected for agreement by chance. The kappa-statistic as proposed by Cohen (1960) is a coefficient to evaluate the agreement among several raters. We have the observations of two raters and assume that both raters classify statistically independent. The first mention of a kappa-like statistic is attributed to Galton (1892), see Smeeton (1985).
The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

In broad terms a kappa below 0.2 indicates poor agreement and a kappa above 0.8 indicates very good agreement beyond chance. Given a set of n elements $S = \{O_1, O_2, \ldots, O_n\}$ and two partitions of S to compare, $X = \{x_1, x_2, \ldots, x_r\}$ and $Y = \{y_1, y_2, \ldots, y_s\}$, The Rand index, R, is:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Intuitively, $a + b$ can be considered as the number of agreements between X and Y and $c + d$ as the number of disagreements between X and Y. The crand index is the Rand index corrected for agreement by chance. Fig.3, Fig.4 and Table1 shows the various performance measures.

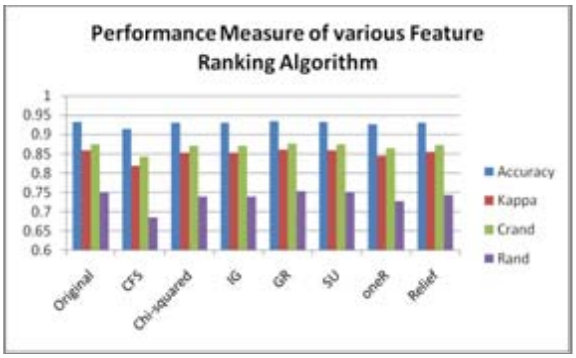


Fig.3. Averaged Performance measures of various Feature Ranking methods.

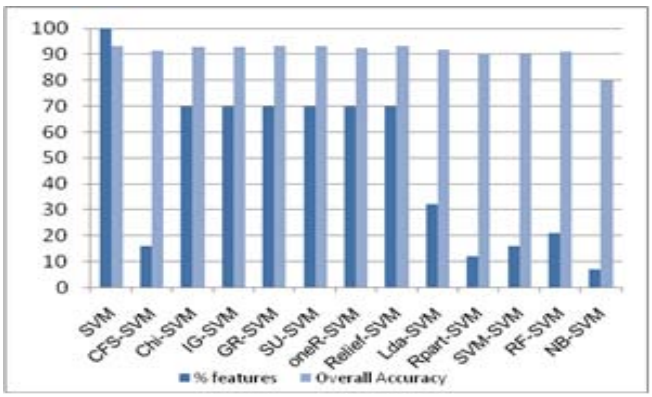


Fig.4. Averaged Classification accuracy of various Filter and Wrapper methods.

Methods	Feature %	Accuracy
SVM	100	93.27
CFS-SVM	16	91.44
Chi-SVM	70	93.00
IG-SVM	70	93.00
GR-SVM	70	93.39
SU-SVM	70	93.33
oneR-SVM	70	92.65
Relief-SVM	70	93.15
Lda-SVM	32	91.90
Rpart-SVM	12	90.51
SVM-SVM	16	89.95
RF-SVM	21	91.23
NB-SVM	7	80.00

Table1: A comparison of Feature Percent and Accuracy

IX. Conclusion

In this paper, we experiment several FS strategies to work on the spam e-mail data set. On the whole, the strategies with RBF kernel are better than the ones without it. In our evaluation, we test how the implemented FS can affect (i.e. improve) the accuracy of Support vector machine classifiers by performing FS. The results show that filter method CFS, Chi-squared, GR, ReliefF, SU, IG, oneR, enabled the classifiers to achieve the highest increase in classification accuracy on the average while reducing the number of unnecessary attributes. The primary purpose of FS is to

reduce the dimensionality to decrease the computation time. This is particularly important concerning text categorization where the high dimensionality of the feature space is a problem. In many cases the number of features is in the tens of thousands. Then it is highly desirable to reduce this number, preferably without any loss in accuracy. The reason for using these five FS methods CFS, LDA, RF, Rpart and NB among twelve FS methods in this study is that they all have shown good performance.

The experiments have shown that in many cases CFS gives results that are comparable or better than the wrapper, Because CFS make use of all the training data at once. The number of features selected by the wrapper using CFS is very Less is very faster than the wrapper, by more than an order of magnitude, which allows it to be applied to large size of the datasets than the wrapper.

X. Acknowledgement

The authors thank many people who have contributed to the R Package; in particular, acknowledgement to all contributors, R statistics, tools and code for their invaluable efforts.

References Références Referencias

1. Hall, M. A., Smith, L. A., 1997, Feature Subset Selection: A Correlation Based Filter Approach, International Conference on Neural Information Processing and Intelligent Information Systems, Springer, p855-858.
2. Kira, K., and Rendell, L. A. The feature selection problem: Traditional methods and a new algorithm. In Proceedings of the AAAI-92 (1992), AAAI Press, pp. 129-134.
3. Kira, K., and Rendell, L. A. A practical approach to feature selection. In The 9th International Conference on Machine Learning (1992), Morgan Kaufmann, pp. 249-256.
4. John, G. H., Kohavi, R., and Pflegger, K. Irrelevant features and the subset selection problem. In Machine learning: Proceedings of the Eleventh International Conference (1994), Morgan Kaufmann, pp. 121-129.
5. Sahami, M., Dumais S., Heckerman D., and Horvitz, E. (1998), A Bayesian approach to filtering junk e-mail. Learning for Text Categorization, Papers from the AAAI Workshop, Madison Wisconsin, pp. 55–62. AAAI Technical Report WS-98-05.
6. A.M. Mesleh, CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System, Proceedings of the 2nd International Conference on Software and Data Technologies, (Knowledge Engineering), Vol. 1, Barcelona, Spain, July, 22—25, 2007, pp. 235-240.
7. Ghiselli E.E. Theory of Psychological Measure_ment, McGraw_Hill.
8. J.R. Quinlan, Induction of decision trees, Machine Learning 1, 81-106, 1986.
9. Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Networks, 5(4):537–550.
10. Press, W. H., Flannery, B. P., Teukolsky, S. A., &Vetterling, W. T. (1988). Numerical recipes in C Cambridge University Press, Cambridge.
11. Holte, R.C. (1993) “Very simple classification rules perform well on most commonly used datasets.” Machine Learning, Vol. 11, 63–91.
12. Ginsberg, M. L 1993, Essentials of Artificial Intelligence, Morgan Kaufmann.
13. Duchene and S. Leclercq, "An Optimal Transformation for Discriminant Principal Component Analysis," IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 10, No 6, November 1988.
14. Breiman, L., 1998, “Arcing classifiers. Annals of Statistics, 26(3):801– 849”.
15. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984, Classification and Regression Trees. Wadsworth International, Belmont, Ca.
16. Therneau TM, Atkinson EJ (1997). \An Introduction to Recursive Partitioning Using the rpart Routine." Technical Report 61, Section of Biostatistics, Mayo Clinic, Rochester,URL http://www.mayo.edu/hsr/techrpt/61.pdf.
17. I. Rish, An empirical study of the naive Bayes classifier, IJCAI 2001, Workshop on Empirical Methods in Artificial Intelligence.
18. Vapnik V N., 1995, The nature of statistical learning theory. New York, Springer.
19. Vladimir N. Vapnik, The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995, 187 pp.
20. Yang, Y., Pedersen, J.O., A Comparative Study on Feature Selection in Text Categorization, Proc. of the 14th International Conference on Machine Learning ICML97, pp. 412---420, 1997.
21. Androutsopoulos, I., Koutsias, J.: An Evaluation of Naive Bayesian Networks. In:Machine Learning in the New Information Age. Barcelona Spain (2000) 9-17.
22. Cristianini, N., and Shawe-Taylor, J., 2000, “An introduction to support vector machines. Cambridge, UK: Cambridge University Press”.
23. Cristianini, N., and Shawe-Taylor, J., 2003”. Support Vector and Kernel Methods,

Intelligent Data Analysis: An Introduction Springer – Verlag".
24. Schölkopf, B., Burges, C.J.C., and Smola, A.J., (Eds.), 1998," Advances in Kernel Methods: Support Vector Learning, MIT Press.
25. Smola, A.J., and Scholkopf, B., Learning with kernels: Support Vector Machines, regularization, optimization, and beyond, Cambridge, MA: MIT press".
26. Burges, C.J.C., 1998," A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121–167".
27. Karatzoglou , A., Smola, A., Hornik, K,, Zeileis,A.,2005,"kernlab – Kernel Methods." R package, Version 0.6-2. Available from http://cran.R-project.org.
28. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004, " kernlab – An S4 Package for Kernel Methods in R." Journal of Statistical Software,11(9). URL http://www.jstatsoft.org/v11/io9/".
29. Chih-Chung Chang., Chih-Jen Lin.,, 2001, "Libsvm: a library forsupport vector machines.http://www.csie.ntu.edu.tw/~cjlin/libsvm".
30. R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0,URLhttp://www.R-project.org/.
31. Leisch,F.,Dimitriadou,E.,2001,"mlbench—A Collection for Artificial and Real-world Machine Learning Benchmarking Problems."R package, version 0.5-6. Available from http://CRAN.R-project.org.
32. Hettich, S., Blake, C. L., and Merz, C. J., 1998," UCI repository of Machine learning databases, Department of Information and Computer Science, University of California, Irvine, CA",.http://www.ics.uci.edu/~mlearn/MLRepository.html"
33. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A (2005), e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, Version 1.5-11, URL http://CRAN.R-project.org/.
34. Dash, M., and Liu, H., 1997, Feature selection for classification. Intelligent Data Analysis: An International Journal", Vol. 1(3), pp. 131- 156.
35. Domingos, P., & Pazzani, M. (1996). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier, Proceedings of the International Conference on Machine Learning.

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

REBEE – Reusability Based Effort Estimation Technique using Dynamic Neural Network

By Jyoti Mahajan, Devanand, Kashyap Dhruve
University of Jammu

Abstract- Software Effort Estimation has been researched for over 25 years but until today no real effective model could be designed that could efficiently gauge the effort required for heterogeneous project data. Reusability factors of software development have been used to design a new effort estimation model called REBEE. This encompasses the usage of Fuzzy Logic and Dynamic Neural Networks. The experimental evaluation of the model depicts efficient effort estimation over varied project types.

Keywords: *Software Effort Estimation, Reusability, Dynamic Neural Networks, Fuzzy Logic, REBEE.*

GJCST Classification: *D.2.9, F.1.1*

Strictly as per the compliance and regulations of:

REBEE – Reusability Based Effort Estimation Technique using Dynamic Neural Network

Jyoti Mahajan[α], Devanand[Ω], Kashyap Dhruve[β]

Abstract- **Software Effort Estimation has been researched for over 25 years but until today no real effective model could be designed that could efficiently gauge the effort required for heterogeneous project data. Reusability factors of software development have been used to design a new effort estimation model called REBEE. This encompasses the usage of Fuzzy Logic and Dynamic Neural Networks. The experimental evaluation of the model depicts efficient effort estimation over varied project types.**

Keywords- Software Effort Estimation, Reusability, Dynamic Neural Networks, Fuzzy Logic, REBEE.

I. Introduction

For over two decades now researchers have developed varied methods to estimate the software effort required to complete a software development project but till date no conclusive method has evolved. Software Effort Estimation is vital to arrive at development effort required for Project Management. Effective software effort estimation techniques not only enable fruitful resource allocation, resource scheduling, risk assessment but also assist in project monitoring. Effective Effort Estimation techniques are useful for fiscal estimates and delivery timelines too.

Effort Estimation techniques could be broadly classified as

- Parametric Effort Estimation Techniques

  The Parametric effort estimation technique assumes the Software Development Cycle to be completely sequential and automated. The software effort required is calculated based on a set of parameters. Once the effort estimated is derived the stringent processes are followed to meet the timelines.

The parametric effort estimation technique neglects the variance in the learning, execution and programming capabilities of every individual involved in the project. Technology dynamics also cannot be analyzed using the parametric effort estimation model.

About[α]- Computer Engineering Department, Govt. College of Engineering & Technology, Jammu, India
E-mail : jyoti_1972@sify.com
About[Ω]- Department of Computer Science & IT, University of Jammu, India
E-mail : padhadevanand@yahoo.co.in
About[β]- Planet-i-Technologies, Bangalore, India
E-mail : kashyap@vardhanatech.com

- Top down approach based effort estimation techniques.

The top down approach considers the entire project as a whole and defragments it into various smaller units. Effort Estimations are carried out for the fragments mainly based on expert judgment. The top down approach does not account for technological changes, future uncertainties and risk mitigation techniques.

- Bottom up approach based effort estimation techniques

The bottom up approach considers using smaller project modules for the construction of the entire project. The entire effort estimation is a summation of the efforts involved in the smaller modules. This approach has drawbacks similar to the top down approach discussed above.

- Analogy Based Approach for effort estimation

The analogy based approach for effort estimation could be considered effective as it is capable to handle dynamics of technological platform transformations, varied human behavior, risk mitigation techniques etc. Prior knowledge about similar projects leads to effective estimation in this approach.

The software industry has experienced tremendous growth over the past decade. Currently Software development contracts are awarded to organizations having a previous experience in handling similar project or related projects. This is done in order to assure quality, reliability, financial security and most importantly timely delivery. Surveys conducted have found that many of these projects fail [1]. Some of the projects encounter effort overruns or schedule overruns sometimes both, due to un-appropriate estimation technique used [2] though prior experience and knowledge is available.

This paper proposes a Reusability Based Effort Estimation Technique (REBEE) to address the issue. REBEE embodies a Neuro-Fuzzy engine for estimation. To handle the dynamics of all the parameters involved in Software Effort Estimation Dynamic Neural Networks is used.

The manuscript is organized as follows. Section 2 describes the existing Effort Estimations Techniques used. Section 3 describes the prominence of reusability and its effects. Dynamic Neural Networks used in the REBEE is discussed in Section 4. Section 5 discusses the REBEE model. Section 6 provides a practical

application example. The paper conclusions with summary and recommendations are provided in Section7.

It can be stated that

Delivery Time α Effort Required …….. (1)

II. Background

Software Effort Estimation form it emergence has been achieved using various methodologies. COCOMO [3] and COCOMO 2.0[4], DELPHI [5], Function Point [6], Planning Poker[7], Use Case Point[8],Expert judgment [9], IBM – FSD [10] based estimation techniques are commonly used. All these models established have drawbacks leading to gross error of estimation. Researchers have used additional techniques along with these to achieve improved efficiency. COCOMO with effort adjustment factor [11] provided about 30% improvement in effort variance. COCOMO used with fuzzy logic, trapezoidal function and Gaussian functions showed improved performance [12].As the regularly used estimation techniques failed to provide consistency when tested against several cases. Multiple software effort estimation techniques were integrated and their combinations were linearly weighed providing better results [13].

A clustering approach bundled with Support Vector regression was found to provide good estimation accuracy [14]. The Mantel's correlation randomization test named Analogy-X greatly improved the estimation algorithm performance [15].The after effects of Schedule and Budget pressure on Effort Estimation and the development cycle time has been closely studied[16]. Chronological Splits have to be carefully assigned for effective training and testing purposes [17]. Judgment Based Systems[18] and recommendation based[19] effort estimation techniques provided satisfactory effort predictions. Global Software Developments being executed at diverse locations worldwide also encounter inaccurate estimation techniques [20].

With numerous efforts estimation techniques available and none providing homogenous results it becomes difficult to decide whether formal models like COCOMO etc or human judgment based systems could be considered ideal for developing effort estimation models [21]. A combination of judgment based and formal based models could be considered as a ideal solution. REBEE proposed in this paper is a combination of formal models developed using dynamic neural networks in addition to judgment based reusability matrices which is discussed in the next section.

III. Reusablity in Effort Estimation

Effort Estimation is a prominent feature of the software development cycle. As studied none of the existing models could effectively predict the software effort for varied types of software projects. The software industry is matured and experience in handling similar kind of projects has provided for additional software development projects of similar nature being offered to organizations. These organizations face a mammoth task of effort estimation. Conventional formal models do not predict the effort accurately as they have not incorporated the Reusability Factor within them. Most of the forms of software development would consist of reusable components like dynamic link libraries, functions, test cases ,web services , etc. Reusability of codes is analyzed seriously by software development organizations that are also considered for appraisals of programmers [22]. Reusable codes would allow software development houses to cut costs [23], reduce effort and maximize profits.

Reusability is a very important factor being analyzed by researchers. Reusability based cost estimation models have been analyzed and the incorporation of the reusable weights into the existing COSYSMO let to a new model called COSYSMO reuse extension[24]. The conventional taguchi model incorporated with reusability exhibited efficient effort estimation results [25]. Reusability incorporation with COCOMO81, COCOMO2 [26] and COCOMO [11] has been analyzed to understand the model performance. While developing REEBEE we considered the importance as well as the ill effects of reusability in the development of the model [27].

IV. Use of Dynamic Neural Networks in Rebee

Static Neural Networks possess learning and adaptive capabilities only for static input output relationships. But when we consider non linear mapping functions that exist in the matrices or parameters used for software effort estimation static neural networks would not be capable of handling the dynamics efficiently. REBEE is developed using dynamic neural networks (DNN). DNN are capable of providing instantaneous outputs for linear or non linear mapping functions that are required to effectively estimate the software effort required. A dynamic neuron unit (DNU) is considered as a basic computing block of the DNN.

A simple DNN *i* is as shown in Fig. 1.

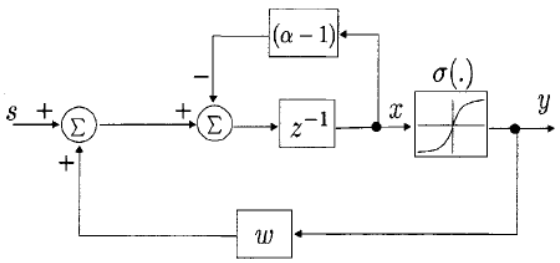


Fig.1. A simple i^{th} DNU

A simple structure of the DNN is shown in Fig. 2. given below.

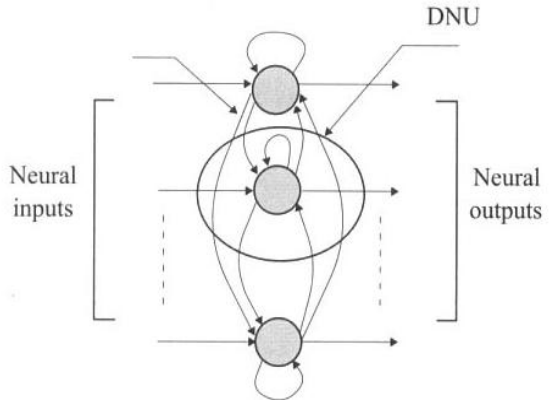


Fig.2. A Simple DNN Structure

Given a finite length discrete time sequence $x_d(k), k = 1,2,\ldots\ldots,N$, we wish to design a discrete-time temporal learning algorithm such that the state of the following discrete-time dynamic neural unit (DT-DNU)

$$x(k+1) = -(a-1)x(k) + \sum_{i=1}^{n} a_i \sigma b_i x(k) + c_i) + s(k)$$

$$= -(a-1)x(k) + f(x(k),w) + (k)$$

$$= -(a-1)\,x\,(k) + a^T\sigma(bx(k) + c) + (k)$$

Will asymptotically track the sequence $x_{d(k)}$.

$$f(x,w) = \sum_{i=1}^{n} a_i \sigma(b_i x + c) = a^T \sigma\,(bx+c)$$

In this case, an error index with quadratic form is defined by

$$E\,(k) = \tfrac{1}{2}(x_d\,(N)-x+(N))^2 + \tfrac{1}{2}\sum_{i=1}^{n-1} [x_d\,(k)-x(k)]^2$$

$$= \tfrac{1}{2}e^2\,(N) + \tfrac{1}{2}\sum_{i=1}^{n-1} e^2\,(k)$$

Where $e(k) = x_{d\,(k)-x(k)}$ and $e(N) = x_{d(N)-x(N)}$.

Using the discrete-time variational principle, a discrete-time lagrangian is defined by

$$\Phi = \tfrac{1}{2}(x_d\,(N)-x(N))^2 + \sum_{k=0}^{N-1} \{\tfrac{1}{2}x_d\,(k)-x(k)^2$$

$$-z(k+1)[x(k+1)+(a-1)x(k)-f(x(k),w)-s(k)]\}$$

$$= \tfrac{1}{2}e^2(N) + \sum_{k=0}^{N-1}\{\tfrac{1}{2}e^2(N) - z(k+1)[x(k+1)$$

$$+ (\alpha-1)x(k) - f(x(k),\boldsymbol{w}) - s(k)]\}$$

The reason that the disrete timev$(k+1)$ is associated with the Lagrange multiplier is due to the unfuziness of the final condition, as will be clear in the following discussion.

Similar to the method used for the continuous-time case, the first variation $\Phi$of may be represented as

$$\delta\Phi = e(N)\delta x(N) + \sum_{k=0}^{N-1}\{e(k)\delta x(k)$$

$$- z(k+1)[\delta x(k+1) + (\alpha-1)\delta x(k) + x(k)\delta\alpha$$

$$- f_x(x(k),\boldsymbol{w})\delta x(k) - \boldsymbol{f}_w(x(k),\boldsymbol{w})^T\delta\boldsymbol{w}]\}$$

$$= e(N)\delta x(N) + \sum_{k=0}^{N-1}\{[e(k) - (\alpha-1)z(k+1)$$

$$+ f_x(x(k),\boldsymbol{w})z(k+1)]\delta x(k) - z(k+1)\delta x(k+1)$$

$$- z(k+1)x(k)\delta\alpha + z(k+1)(\boldsymbol{f}_w(x(k),\boldsymbol{w}))^T\delta\boldsymbol{w}\}$$

Let the Lagrange multiplier $z(k)$satisfy,

$$z(k) = e(k) + [f_x(x(k),\boldsymbol{w}) - (\alpha-1)]z(k+1)$$

Or

$$z(k+1) = \frac{z(k) - e(k)}{f_x(x(k),\boldsymbol{w}) - (\alpha-1)}$$

Then

$$\delta\Phi = e(N)\delta x(N) + \sum_{k=0}^{N-1}[z(k)\delta x(k) - z(k+1)\delta x(k+1)$$

$$- z(k+1)x(k)\delta\alpha + z(k+1)(\boldsymbol{f}_w(x(k),\boldsymbol{w}))^T\delta\boldsymbol{w}]$$

$$= z(0)\delta x(0) - [z(N) - e(N)]\delta x(N)$$

$$+ \sum_{k=0}^{N-1}[-z(k+1)x(k)\delta\alpha + z(k+1)(\boldsymbol{f}_w(x(k),\boldsymbol{w}))^T\delta\boldsymbol{w}]$$

Since the initial value $x(0)$ does not depend on the parameters $\delta x(0) = 0$. If we choose additionally the final condition of the Lagrange multiplier

$$z(N) = e(N)$$

Then,

$$\delta\Phi = \sum_{k=0}^{N-1}[-z(k+1)x(k)\delta\alpha + z(k+1)(\boldsymbol{f}_w(x(k),\boldsymbol{w}))^T\delta\boldsymbol{w}]$$

$$= \left(\sum_{k=0}^{N-1} -z(k+1)x(k)\right)\delta\alpha + \left(\sum_{k=0}^{N-1} z(k+1)(\boldsymbol{f}_w(x(k),\boldsymbol{w}))^T\right)\delta\boldsymbol{w}$$

Therefore, the partial derivatives of the error index with respect to the parameters are given by

$$\frac{\partial E}{\partial \alpha} = -\sum_{k=0}^{N-1} z(k+1)x(k)$$

$$\frac{\partial E}{\partial \boldsymbol{w}} = \sum_{k=0}^{N-1} z(k+1)\boldsymbol{f}_w(x(k),\boldsymbol{w})$$

And the incremental terms of the parameters are

$$\Delta\alpha(k) = -\eta_\alpha\frac{\partial E}{\partial \alpha} = \eta_\alpha\sum_{k=0}^{N-1} z(k+1)x(k)$$

$$\Delta\boldsymbol{w}(k) = -\eta_w\frac{\partial E}{\partial \boldsymbol{w}} = -\eta_w\sum_{k=0}^{N-1} z(k+1)\boldsymbol{f}_w(x(k),\boldsymbol{w})$$

That is, the updating equations are obtained as

$$\alpha(k+1) = \alpha(k) + \eta_\alpha\sum_{k=0}^{N-1} z(k+1)x(k)$$

$$\boldsymbol{w}(k+1) = \boldsymbol{w}(k) - \eta_w\sum_{k=0}^{N-1} z(k+1)\boldsymbol{f}_w(x(k),\boldsymbol{w})$$

The Learning algorithm given above for such a sequence learning problem consists of a discrete-time two-point boundary-value problem that can be solved, in general, by reiterative technique. Here, the initial condition $x(0)$ of the state is known, and the final condition $z(N)$ of the Lagrange multiplier is a linear function of the unknown final condition $x(N)$ of the state.

From the above discussion we can analyze the behavior and the working of the DNN trained using the Back Propagation Algorithm. A sigmoid function is used as the activation function.

V. REBEE

In this section we would discuss about the functional properties of REEBEE. The working of REEBEE could be understood through the following steps. Let us consider the dataset provided to us represented by *D* matrice. *D* is a *m×n* matrix represented as

$$D_{(m\times n)} = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mn} \end{bmatrix}$$

Where p are the Parameters of the Data Set.

Using Fuzzy Rules we now need to derive the reusability matrix represented as R of the parameters presented in the dataset. The matrix R consists of the effort taken to achieve the reusable part of the parameters p. The reusability matrix is a i×j matrix represented as

$$R_{(i\times j)} = \begin{bmatrix} r_{11} & \cdots & r_{1j} \\ \vdots & \ddots & \vdots \\ r_{i1} & \cdots & r_{ij} \end{bmatrix}$$

Let fzp(x) represent the fuzzy rule such that

$$R_{(i\times j)} = f_{zp}\left(D_{(m\times n)}\right)$$

The fuzzy matrix is provided to the DNN of the REEBEE to obtain the estimated effort matrix E. E also is a m×n represented as. p(m×n) corresponds to the effort estimated in man hours / man months

$$E_{(m\times n)} = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mn} \end{bmatrix}$$

VI. A Practical Application Example

To evaluate the robustness of REBEE we conducted experimental evaluations on 2 data sets obtained from a $5 billion per year international technology firm. The datasets contained data of a testing and an enterprise application development project. The data was presented in spreadsheet formats. To interface these data sets we incorporated a "Import Wizard" in REEBEE which also embodied the Fuzzy Rules to derive the reusability matrix.

On Importing the dataset into REEBEE the Import Wizard Provided the Reusability Matrix to the DNN. The DNN was trained using the reusability matrix. The learning Rate used was set to 0.01, Number of Iterations were 10,000. The Sigmoid Function was used as the activation function. The experimental results obtained are as shown graphically in the figure below.

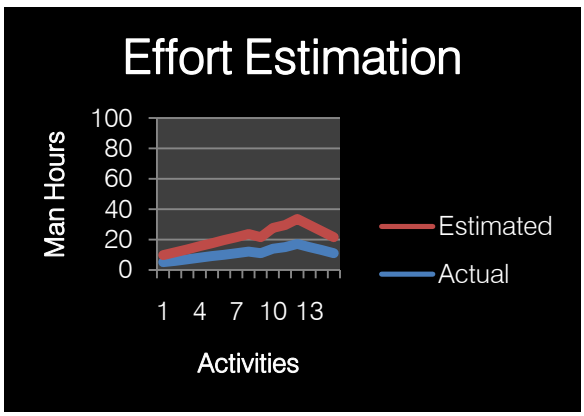


Fig.3. Effort Estimated and Actual Effort Time for Software Testing Project

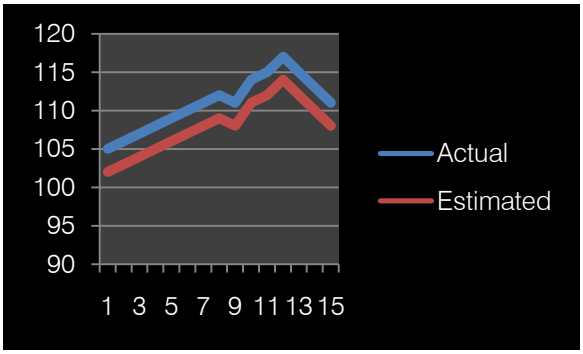


Fig.4. Effort Estimated and Actual Effort Time for Software Testing Project

From Fig. 3. and Fig. 4. we could conclude that the REBEE estimation is very close to the actual effort.

VII. Conclusion and Future Work

Effort Estimation is a critical operation of the Software Development Cycle. The existing estimation techniques do not provide similar estimation results for varied projects. Reusability factor is predominant in the current Software Development Cycle. The research work presented here introduced REEBEE, a reusability based effort estimation technique. REBEE is a neuro-fuzzy system which embodies fuzzy rules to derive reusability matrices and dynamic neural networks for effort estimation based on the reusability matrix. The experimental study conducted on data sets of 2 different projects showed impressive differences concluding that the REEBEE is capable of Effective Software Effort Estimation Techniques.

In the future we would like to further investigate the performance of REBEE over different project types and study its responses.

References Références Referencias

1. C. E. L. Peixoto, J. L. N. Audy, R. Prikladnicki, "Effort Estimation in Global Software Development Projects: Preliminary Results from a Survey," in Proc. 2010 *5th IEEE International Conference on Global Software Engineering, pp.123-127.*
2. K. Molkken, M. Jorgensen, "A Review of Surveys on Software Effort Estimation," Proc. 2003 *International Symposium on Empirical Software Engineering (ISESE'03), pp. 223.*
3. B.W. Boehm, W.W. Royce, Le COCOMO Ada, *Genie logiciel & Systemes experts*, 1989
4. B.W. Boehm, et al., "Cost Models for Future Software Life Cycle Processes: COCOMO2.0", *Annals of Software Engineering on Software Process and Product Measurement, Amsterdam,* 1995.
5. Website http://www.stellman-greene.com/aspm/images/ch03.pdf.
6. Meli, R., L. Santillo, "Function point estimation methods: a comparative overview", in Proc. 1999 *The European Software Measurement Conference – Amsterdam,* October 6-8.
7. http://www.mountaingoatsoftware.com/system/presentation/file/51/bayXP_070320_PlanningAgileProjects.pdf
8. S. Nageswaran "Test effort estimation using use case points" in *14th International Internet Software Quality Week 2001, San Francisco, California, USA,* June 2001.
9. M. Jørgensen, "Practical Guidelines for Expert-Judgment-Based Software Effort Estimation," *IEEE Software*, vol. 22, no. 3, pp. 57-63, May/June 2005.
10. C.E. Walston, A.P. Felix, "A method of programming measurement and estimation," *IBM Systems Journal*, vol. 16, no.1, 1997.
11. M.J. Basavaraj, K.C Shet, "Empirical validation of Software development effort multipliers of Intermediate COCOMO Model" *Journal of Software*, vol. 3, vo. 5, pp 65 MAY 2008.
12. C.S. Reddy, KVSVN Raju, "An Improved Fuzzy Approach for COCOMO's Effort Estimation using Gaussian Membership Function". *Journal of Software*, vol. 4, no. 5, pp. 452-459, 2009.
13. C.J. Hsu, N.U. Rodas, C.Y. Huang, K.L. Peng, "A Study of Improving the Accuracy of Software Effort estimation Using Linearly Weighted Combinations," in Proc. 2010 *IEEE 34th Annual Computer Software and Applications Conference Workshops*, pp.98-103.
14. E. Kocaguneli, A. Tosun, A. Bener. "AI-Based Models for Software Effort Estimation" in Proc. *36th EUROMICRO Conference on Software Engineering and Advanced Applications*, pp.323-326.
15. J.W. Keung, B.A. Kitchenham, D.R. Jeffery, "Analogy-X: Providing statistical Inference to Analogy-Based Software Cost Estimation," *IEEE Transactions on Software Engineering*, vol. 34, no. 4, pp. 471-484, July/Aug. 2008.
16. N. Nan, D.E. Harter "Impact of Budget and Schedule Pressure on Software Development Cycle Time and Effort" *IEEE Transactions on Software Engineering*, vol. 35, no. 5, pp 624-637, 2009.
17. C. Lokan, E. Mendes "Investigating the Use of Chronological Split for Software Effort Estimation" *IET*-Software, vol. 3, no. 5, pp 422-434, October 2009.
18. S. Grimstad, M. Jorgensen "Preliminary study of sequence effects in judgment-based software development work-effort estimation", *IET - Special Issue (EASE)* vol. 3 no. 5, pp 435-441, 2009.

19. B. Peischl, M. Nica, M. Zanker "Recommending effort estimation methods for software project management", in Proc. IEEE/*WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, vol. 03, pp 77-80, 2009.
20. C.E.L. Peixoto, J.L.N. Audy, R. Prikladnicki, "Effort Estimation in Global Software Development Projects: Preliminary Results from a Survey," in Proc. 2010 *5th IEEE International Conference on Global Software Engineering, ICGSE*, pp.123-127.
21. M. Jorgensen, B.W. Boehm, "Software Development Effort Estimation: Formal Models or Expert Judgment" *IEEE Software,* vol. 26, no. 2, pp. 14-19, Mar./Apr. 2009.
22. P.S. Sandhu, H. Singh, "Automatic Reusability Appraisal of Software Components using Neuro-Fuzzy Approach", *International Journal of Information Technology*, vol. 3, no. 3, pp. 209-214, 2006.
23. P.S. Sandhu, H. Kaur, A. Singh , "Modeling of Reusability of Object Oriented Software
1. Velmurugan T. and Santhanam T. (2010), "Clustering Mixed Data Points Using Fuzzy C-Means Clustering" retrieved from www.enggjournals.com/ijcse/doc/IJCSE10-02-09-112.pdf.
2. Worobey M, Gemmel M, Teuwen DE (2008) "Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960" retrieved volution.berkeley.edu/evolibrary/news/081101_hivorigins - Cached - Similar.
3. Yang X. and Wand W. (2001) GIS Based Fuzzy C-Means Clustering Analysis of Urban Transit Network Service. The Nanjing City Case Study. Road and Transport Research China.
4. Zadeh L.A. (1965), "Fuzzy sets. Information and Control", Vol.8, pp.338-353.
5. Zain, M.F.M., Islam, M.N. and Basri, H. (2005), "An expert system for mix design of high performance concrete. Advances in engineering software", 36(5): 325 – 337.

System", *Journal of World Academy of Science, Engineering and Technology,* no. 56, pp 162 August 2009.
24. G. Wang, R. Valerdi, J. Fortune, "Reuse in Systems Engineering", *IEEE Systems Journal*, vol. 4, no. 3, pp 376-384, September 2010.
25. P. S. Sandhu, P. Blecharz, H. Singh, "A Taguchi Approach to Investigate Impact of Factors for Reusability of Software Components", *Journal of World Academy of Science, Engineering And Technology*, vol 25, pp 135-140, 2007.
26. CH.V.M.K.Hari, P.V.G.D.P Reddy, J.N.V.R.S Kumar, G.SriRamGanesh. CH.V.M.K. Hari., "Identifying the Importance of Software Reuse in COCOMO81, COCOMOII", *International Journal of Computer Science and Engineering JCSE,* vol. 1 no. 3, pp 142-147, 2009.
27. N. Ozarin, "Lessons Learned on Five Large-Scale System Developments" *IEEE Instrumentation & Measurement Magazine*, nol. 11, Issue- 1, pp 18-23, February 2008.

Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

Up-Down Routing Based Deadlock Free Dynamic Reconfiguration in High Speed Local Area Networks

By Naresh Kumar, Renu Vig, Deepak Bagai
Kurukshtra University

Abstract- Dynamic reconfiguration of high speed switched network is the process of changing from one routing function to another while the network remains in running mode. Current distributed switch-based interconnected systems require high performance, reliability and availability. These systems changes their topologies due to hot expansion of components, link or node activation and deactivation. Therefore, in order to support hard real-time and distributed multimedia applications over a high speed network we need to avoid discarding packets when the topology changes. Thus, a dynamic reconfiguration algorithm updates the routing tables of these interconnected switches according to new changed topology without stopping the traffic. Here, we propose an improved deadlock-free partial progressive reconfiguration (PPR) technique based on UP/DOWN routing algorithm that assigns the directions to various links of high-speed switched networks based on pre-order traversal of computed spanning tree. This improved technique gives better performance as compared to traditional PPR by minimizing the path length of packets to be transmitted. Moreover, the proposed reconfiguration strategy makes the optimize use of all operational links and reduces the traffic congestion in the network. The simulated results are compared with traditional PPR.

Keywords: *Deadlock avoidance, dynamic reconfiguration, UP/DOWN routing, fault tolerance, high-speed networks*

GJCST Classification: *C.2.5*

Strictly as per the compliance and regulations of:

Up-Down Routing Based Deadlock Free Dynamic Reconfiguration in High Speed Local Area Networks

Naresh Kumar[α], Renu Vig[Ω], Deepak Bagai[β]

Abstract- Dynamic reconfiguration of high speed switched network is the process of changing from one routing function to another while the network remains in running mode. Current distributed switch-based interconnected systems require high performance, reliability and availability. These systems changes their topologies due to hot expansion of components, link or node activation and deactivation. Therefore, in order to support hard real-time and distributed multimedia applications over a high speed network we need to avoid discarding packets when the topology changes. Thus, a dynamic reconfiguration algorithm updates the routing tables of these interconnected switches according to new changed topology without stopping the traffic. Here, we propose an improved deadlock-free partial progressive reconfiguration (PPR) technique based on UP/DOWN routing algorithm that assigns the directions to various links of high-speed switched networks based on pre-order traversal of computed spanning tree. This improved technique gives better performance as compared to traditional PPR by minimizing the path length of packets to be transmitted. Moreover, the proposed reconfiguration strategy makes the optimize use of all operational links and reduces the traffic congestion in the network. The simulated results are compared with traditional PPR.

IndexTerms- Deadlock avoidance, dynamic reconfiguration, UP/DOWN routing, fault tolerance, high-speed networks.

I. Introduction

High-performance switch - based interconnected systems require communication between various workstations when there is change in network topologies due to hot expansion of components, failure of links or switches. In past years, many technologies have been proposed to reconfigure the network in case of addition/removal of links/switches without stopping the transmission of packets. Current high-speed switched networks (Myrinet[2,8], Advanced Switching[1], Infiniband[5], Tnet[9]) updates their routing tables when there is change in network topologies due to network components failures and addition of new one. In such cases a dynamic reconfiguration algorithm analyze the new network topology and updates the routing table by replacing the old routing function with the new updated routing function. Although both routing functions are deadlock-free still it may create deadlock during reconfiguration because one of the routing function may take turns that are not allowed in the other routing function by making a cycle in the network.

In the literature, static reconfiguration (designed for Autonet[13,15] and Myrinet[2,8] networks) replaces the routing function from old to new by stopping the network traffic. Hence it creates negative impact on the network service availability and the performance of the overall network degrades.

Many new schemes have been proposed recently to enhance network service availability while the network change over from one routing function to another. These new schemes updates the routing table of switched network during run time when there is any change in the network topologies due to addition/removal of links/nodes and are known as dynamic reconfiguration techniques.(Partial Progressive Reconfiguration (PPR)[3], Skyline[6], Double scheme[10], and Simple Reconfiguration [7]), NetRec[16] and Dynamically scaling Algorithm(DSA) [17]. These schemes are designed for the networks that uses distributed routing. Double Scheme requires extra resources like virtual channels for deadlock handling that occur during transition of old routing function to new routing function. Simple reconfiguration requires a special packet called token to avoid deadlock. In this scheme firstly all packets are transmitted by old routing function, then token and finally the packet transmission is based on new routing function. NetRec was proposed as a dynamic reconfiguration scheme to increase the network availability in the presence of a permanent node fault. It restores the network connectivity by building a tree that spans all immediate neighbours of the faulty node that are still connected to network. The NetRec Scheme[16] requires every switch to maintain information about nodes some number of hops away and is only applicable to wormhole networks. NetRec is

About[α]- University Institute of Engineering and Technology, Kurukshtra University, Kurukshetra (India) (Mobile : +91-94670-12567; fax :+91-1744-238967;
E-mail : naresh_duhan@rediffmail.com
About[Ω]- University Institute of Engineering and Technology , Panjab University , Chandigarh (India).
E-mail :renuvig@hotmail.com
About[β]- Electonics and Electrical Engineering Department, Punjab Engineeering College(Deemed University),Chandigarh.
E-mail: dbagai@yahoo.com

extended to dynamically reconfigure the network for the case of newly joining nodes called Dynamically Scaling Algorithm (DSA). The solution in DSA is based on performing sequence of partial routing table updates, while dropping the user messages in all selected nodes until the restoration is completed. PPR requires a sequence of synchronizing steps to progressively update old forwarding table entries to new ones while ensuring that no cycles form. The PPR [3] approach correct an invalid UP/DOWN graph after change in network topology. Double Scheme[10] uses the concept of virtual channels to avoid deadlock while reconfiguring the network . Simple Reconfiguration [7] uses a packet called a token to avoid deadlock by ensuring that the packet which belongs to old routing function are transmitted first, then the token, and finally packet transmission is based on the new routing function. A management mechanism[11,12] and zero-configuration hierarchical UP/DOWN routing[19] has been discussed for distributed calculation of the new routing function when there is any change in the switched-network topology. In this mechanism the distributed path-computation gives better performance when compared to existing centralized path-computation, still the routing function was updated statically, therefore degrades the performance.

In this research work, we propose to apply a new and very efficient dynamic reconfiguration strategy based on PPR that makes transformation of an invalid UP/DOWN graph (that is due to change in topology) into a valid UP/DOWN graph, while ensuring that there is no cycles in the graph to make it deadlock-free. Moreover, our dynamic reconfiguration does not use any additional resources such as virtual channels or special packet a token. This new proposed scheme gives better performance than PPR by distributing the traffic through optimize use of all links and reducing the congestion on root node of the spanning tree.

The next section discuss the concepts of UP/DOWN routing based PPR scheme and provides background information on previous studies of reconfiguration of UP/DOWN routing networks. Section 3 gives the details of improved dynamic reconfiguration strategy and assigning the directions UP/DOWN to the operational links based on pre-order traversal of spanning tree. In section 4, the performance of the proposed reconfiguration strategy is evaluated. Finally, section 5 concludes this research work and proposes some future work.

II. UP/Down Routing Based Dynamic Reconfiguration (PPR)

PPR scheme is based on deadlock-free UP/DOWN routing algorithm[14] for irregular network topologies . The UP/DOWN routing algorithm is based on a cycle-free assignment of directions to the operational links in the network. For each link in the network, one direction is named up and the other is down. Deadlock avoidance is achieved by using legal routes. A packet never use a link in the up direction after having used one in the down direction. Messages can traverse zero or more links in the up direction, followed by zero or more links in the down direction. Therefore, deadlocks are prevented and cycles in the channel dependency graph[4] are avoided. Major problem of old PPR is the random assignment of UP/DOWN direction to links between two or more nodes at the same label.

A sink node [3] does not have outgoing up-links(a node that is not the source of any link). There are no legal routes between any two sink nodes due to the restrictions imposed by UP/DOWN routing algorithm because each route would require a down to up transition, which is not allowed. So there is always a single sink node in directed acyclic graph based on UP/DOWN routing algorithm. A break node is the source of two or more links. This node breaks the cycles formation to avoid deadlock in directed graph. The graph of Fig 1 includes various break nodes like node j, e, d. A correct graph contains a single sink node for UP/DOWN routing. It includes as many break nodes as necessary to break all cycles for deadlock freedom. The graph shown in Fig 1 is a correct graph. On the other hand, an incorrect graph does not meet the restrictions imposed in the correct graph. An incorrect graph has the absence of a sink node, presence of more than one sink node, or there are cycles in the graph. Fig. 2 tells about an incorrect graph with two sink nodes ‘b’ and ‘c’ after removal of node ‘a’ from Fig. 1 correct graph. A PPR scheme changes the direction of operational links to give a correct graph from an incorrect graph as shown in Fig. 3.In the literature, several algorithms have been proposed for constructing an UP/Down directed graph. Traditional proposals are based on the computation of a spanning tree which is built using a breadth-first search (BFS)[14], a depth-first search(DFS)[13], or a propagation-order spanning tree(POST) [15].

III. An Improved PPR Scheme Based on Pre-Order Traversal of Spanning Tree

Network topology changes due to addition/removal of switches/ links, then our dynamic reconfiguration scheme calculates a new routing function which ensures that packets that belong to the new routing function can not take turns in the old routing function, and vice versa. Therefore, packets of old and new routing functions can unrestrictedly coexist in the network without creating deadlocks. This improved PPR scheme is based on the concept of assigning the UP/DOWN directions to operational links by pre-order

traversal of spanning tree in which a unique label is for each node of the graph. We also present lemmas to support that, when an UP/DOWN graph for the new topology is designed based on pre-order traversal of spanning tree, the routing function is updated without the risk of transient deadlocks.

Definition 1

Assume that two UP/DOWN directed graphs, G1 and G2, represented two network topologies. Then G1 and G2 are corrected.

Fig 4 presents UP/DOWN graph G1 which is correct. Fig. 5 shows again a correct graph G2 after removal of root node 'a' of graph G1 in Fig.4

Lemma1

Assume that an UP/DOWN directed graph G1 based on pre-order traversal of spanning tree is correct. Then, it is always possible to obtain a correct graph G2 from G1 when any node or link is added or removed.

Proof

In a correct UP/Down graph, each possible cycle must contain at least one node with two incoming up-links and at least one node with two outgoing up-links. For each possible

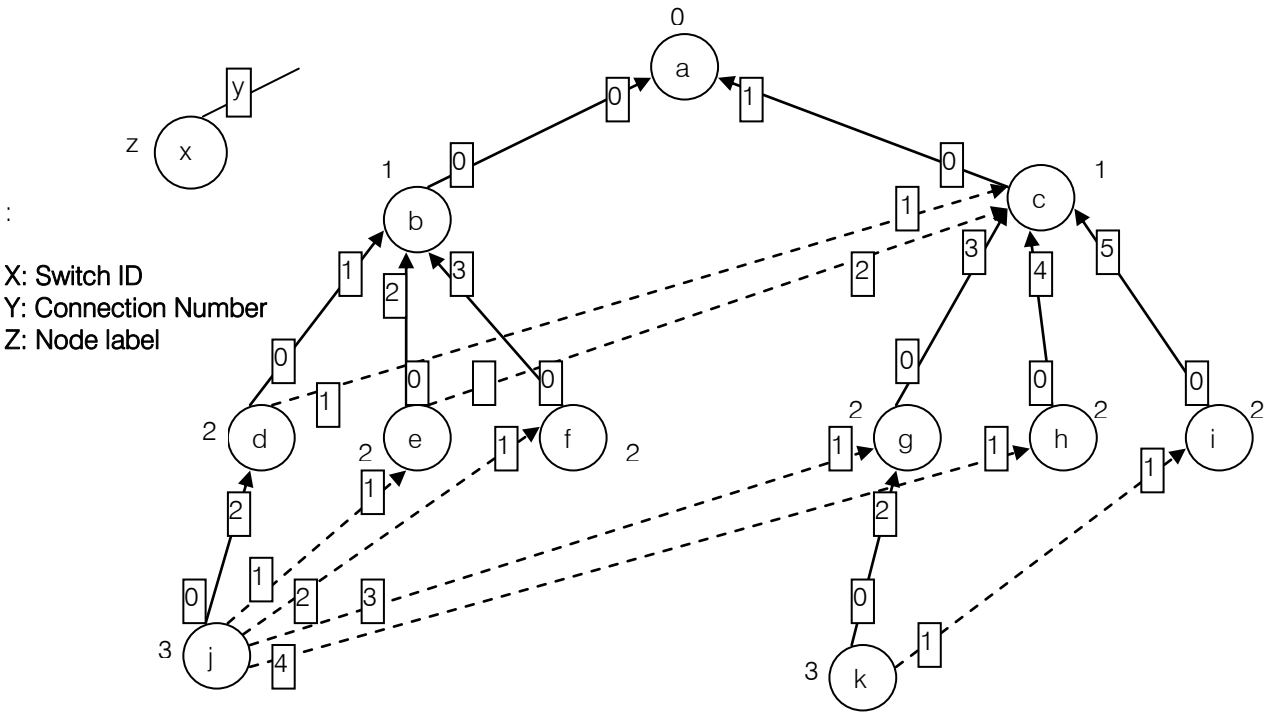


Fig1: Example of UP/DOWN direction assignment in a switched network.

Table 1: Routing table based on UP/DOWN algorithm for Fig 1

Switch	a	b	c	d	e	f	g	h	i	j	k
a	[x]	[0]	[1]	[0,1]	[0,1]	[0]	[1]	[1]	[1]	[0,1]	[1]
b	[0]	[x]	[0]	[1]	[2,0]	[3]	[0]	[0]	[0]	[1,2,3,0]	[0]
c	[0]	[0]	[x]	[1,0]	[2,0]	[0]	[3]	[4]	[5]	[1,2, 3,4,0]	[3,5]
d	[0,1]	[0,1]	[1,0]	[x]	[0,1]	[0]	[1,0]	[1,0]	[1,0]	[2,1,0]	[1,0]
e	[0,2]	[0,2]	[2,0]	[0,2]	[x]	[0]	[2,0]	[2,0]	[2,0]	[1,2,0]	[2,0]
f	[0]	[0]	[0]	[0]	[0]	[x]	[0]	[0]	[0]	[1,0]	[0]
g	[0]	[0]	[0]	[0]	[0]	[0]	[x]	[0]	[1,0]	[2,0]	[0]
h	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[x]	[0]	[1,0]	[0]
i	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[x]	[0]	[1 ,0]
j	[0,1,2, 3,4]	[0,1,2, 3,4]	[1,3,4 ,2,0]	[0,1, 2,3,4]	[1,2,0 ,3,4]	[2,1, 0,3,4]	[3,4, 2,1,0]	[4,3,2, 1,0]	[4,3,2 ,1,0]	[x]	[4,3,2 ,1,0]
k	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[1,0]	[0,1]	[x]

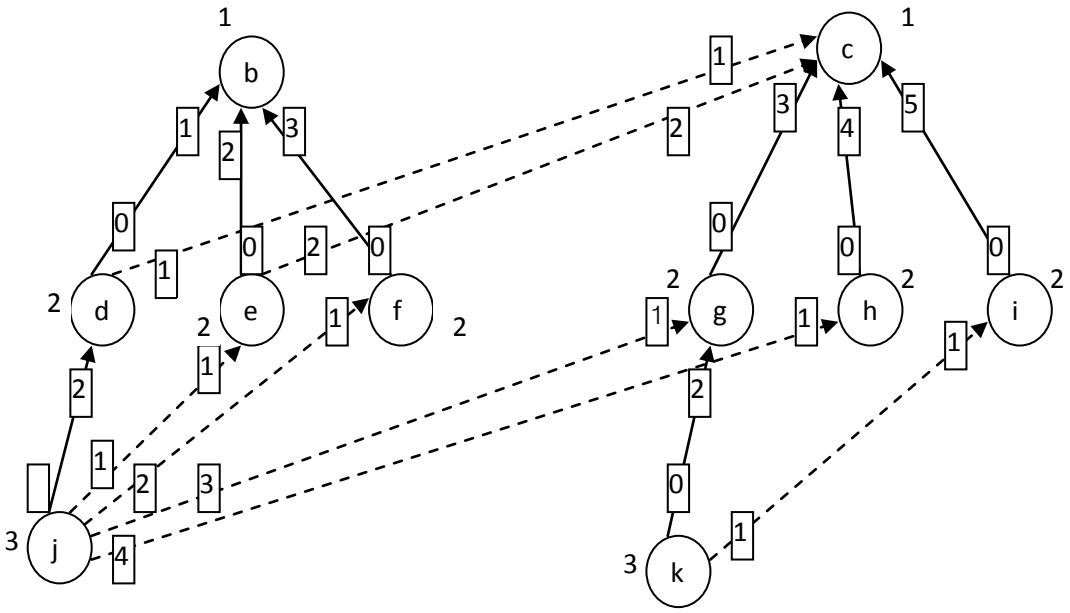


Fig.2 After deactivation of root node 'a' of Fig. 1 (Incorrect graph)

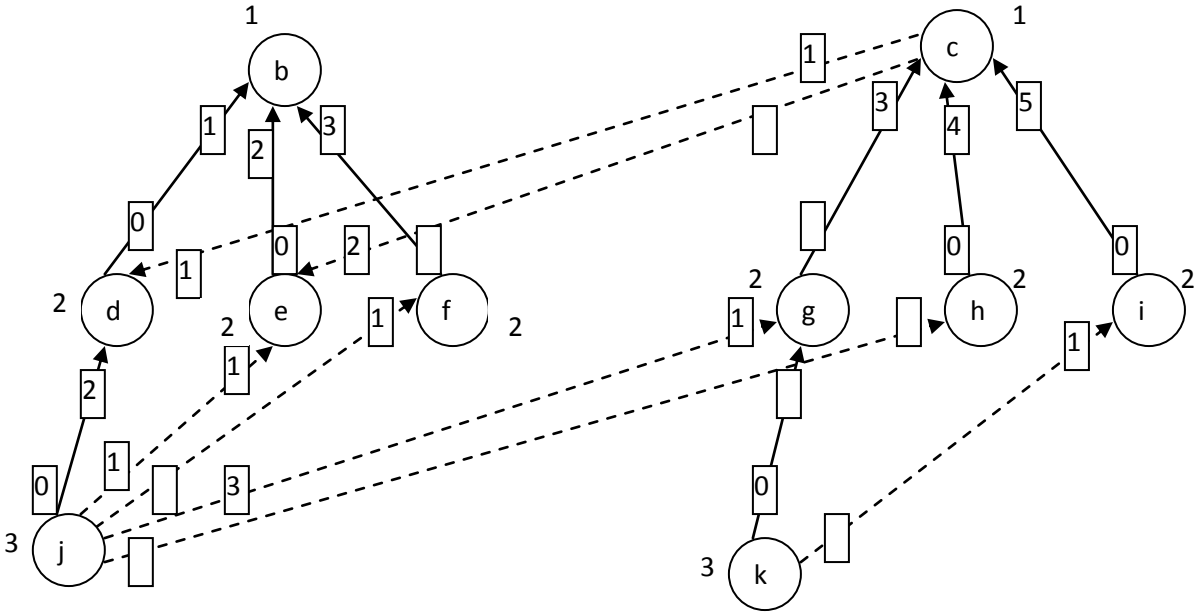


Fig 3. A new directed acyclic correct UP/DOWN graph after deactivation of root node 'a' of Fig. 1 (new correct graph after deactivation)

Table 2: Routing table for Fig 3 based on old PPR after deactivation of node 'a' of Fig. 1

Switch	b	c	d	e	f	g	h	i	j	k
b	[x]	[2]	[1]	[2]	[3]	[1,2]	[1,2]	[1,2]	[1,2,3]	[1,2]
c	[1,2]	[x]	[1,2]	[2,1]	[1,2]	[3]	[4]	[5]	[3,4, 1,2]	[3,5]
d	[0]	[1,0]	[x]	[0]	[0]	[1,0]	[1,0]	[1,0]	[2,1,0]	[1,0]
e	[0]	[2,0]	[0]	[x]	[0]	[2,0]	[2,0]	[2,0]	[1,2,0]	[2,0]
f	[0]	[0]	[0]	[0]	[x]	[0]	[0]	[0]	[1,0]	[0]
g	[0]	[0]	[0]	[0]	[0]	[x]	[0]	[0]	[1,0]	[2,0]
h	[0]	[0]	[0]	[0]	[0]	[0]	[x]	[0]	[1,0]	[0]
i	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[x]	[0]	[1,0]
j	[0,1,2 ,3,4]	[0,1, 3,4,2]	[0.1.2 .3.4]	[1,0,2 ,3,4]	[2,0, 1,3,4]	[3,0, 1,2,4]	[4,0, 1,2,3]	[3,4 ,0,1,2]	[x]	[3,0, 1,2,4]
k	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[1,0]	[1,0]	[x]

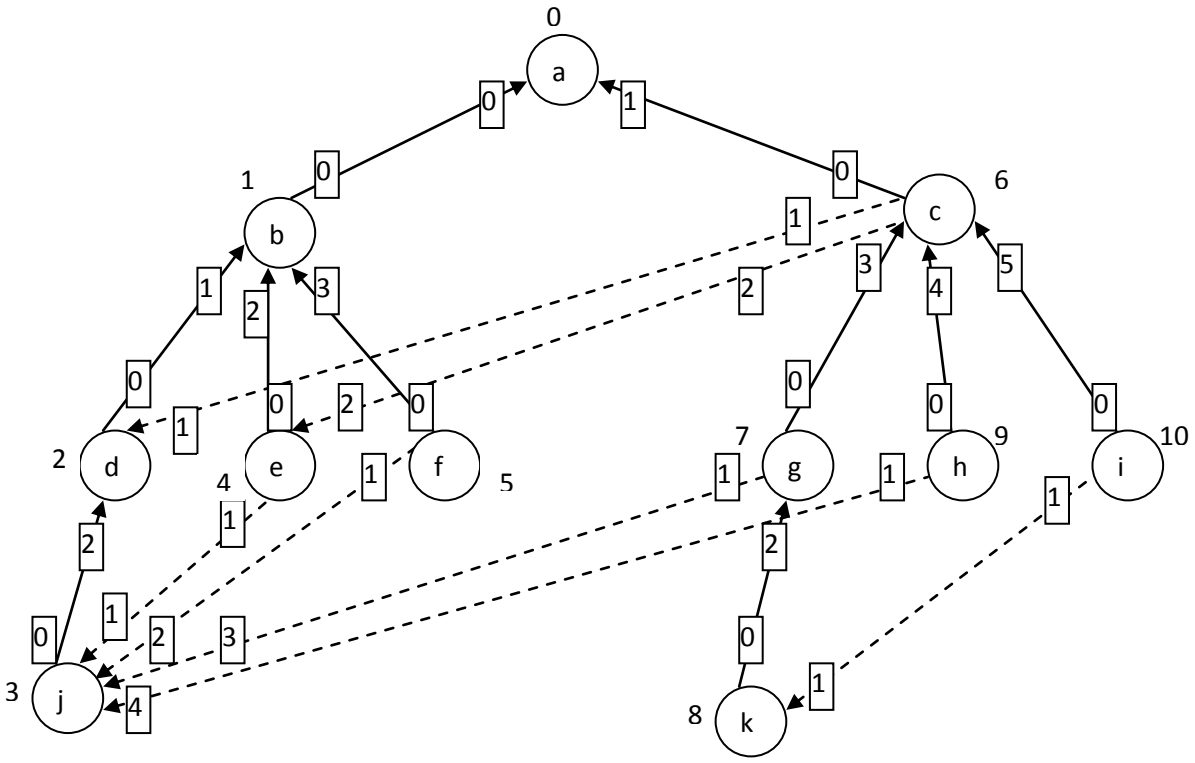


Fig. 4 Pre-order traversal based UP/DOWN routing algorithm (correct graph G1)

Table 3 Routing table for Fig 4 that uses Pre-order traversal based assignment of UP/DOWN direction to various operational links

i

Switch	a	b	c	d	e	f	g	h	i	j	k
a	[x]	[0]	[1,0]	[0]	[0]	[0]	[1,0]	[1,0]	[1,0]	[0]	[1,0]
b	[0]	[x]	[0,2,1]	[1]	[2,1]	[3,1]	[0,1]	[0,1]	[0,1]	[1]	[0,1]
c	[0,1,2]	[0,1]	[x]	[1,0,2]	[2,0,1]	[0,1,2]	[3,2, 1,0]	[4,2, 1,0]	[5,3, 2,1,0]	[2,1,0]	[3,2, 1,0]
d	[0]	[0]	[1,0,2]	[x]	[0,2]	[0,2]	[1,2,0]	[1,2,0]	[1,0,2]	[2]	[1,2,0]
e	[0,1]	[0,1]	[2,0]	[0,1]	[x]	[0,1]	[1,2,0]	[1,2,0]	[2,0,1]	[1,0]	[1,2,0]
f	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[x]	[1,0]	[1,0]	[1,0]	[1,0]	[1,0]
g	[0,1]	[0,1]	[0,1]	[1,0]	[1,0]	[1,0]	[x]	[0,1]	[0,1,2]	[1,0]	[2]
h	[0,1]	[0,1]	[0,1]	[1,0]	[1,0]	[1,0]	[1,0]	[x]	[0,1]	[1,0]	[1]
i	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[x]	[0,1]	[1,0]
j	[0]	[0]	[0]	[0]	[1,0]	[2,0]	[3,0]	[4,0]	[3,1,0]	[x]	[3,1,0]
k	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[1,0]	[0]	[x]

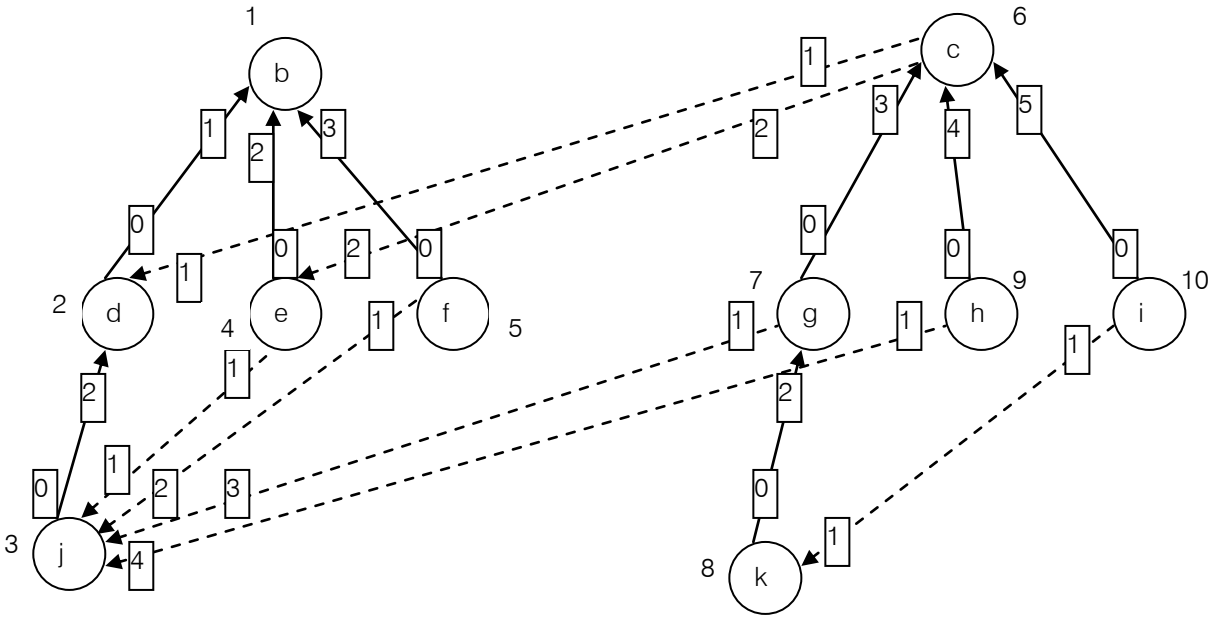


Fig 5 After deactivation of root node 'a' of graph G1 in Fig 4 gives correct graph of G2

Table 4: Routing table for Fig 5 after deactivation of node 'a' of Fig 4

Switch	b	c	d	e	f	g	h	i	j	k
b	[x]	[1,2]	[1]	[2,1]	[3,1]	[1,2]	[1,2]	[1,2]	[1]	[1,2]
c	[1,2]	[x]	[1,2]	[2,1]	[2,1]	[3,2,1]	[4,2,1]	[5,2,1]	[1,2]	[3,2,1]
d	[0]	[1,2,0]	[x]	[0,2]	[0,2]	[2,1,0]	[2,1,0]	[1,0,2]	[2]	[1,2,0]
e	[0,1]	[2,1,0]	[0,1]	[x]	[0,1]	[2,1,0]	[2,1,0]	[2,1,0]	[1,0]	[2,1,0]
f	[0,1]	[0,1]	[0,1]	[0,1]	[x]	[1,0]	[1,0]	[1,0]	[1,0]	[1,0]
g	[0,1]	[0,1]	[0,1]	[0,1]	[1,0]	[x]	[0,1]	[0,2,1]	[1,0]	[2,1,0]
h	[0,1]	[0,1]	[0,1]	[0,1]	[1,0]	[0,1]	[x]	[0,1]	[1,0]	[0,1]
i	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[x]	[1,0]	[1,0]
j	[0]	[1,0]	[0]	[1,0]	[2,0]	[3,0]	[4,0]	[3,1,0]	[x]	[3,1,0]
k	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[1,0]	[0]	[x]

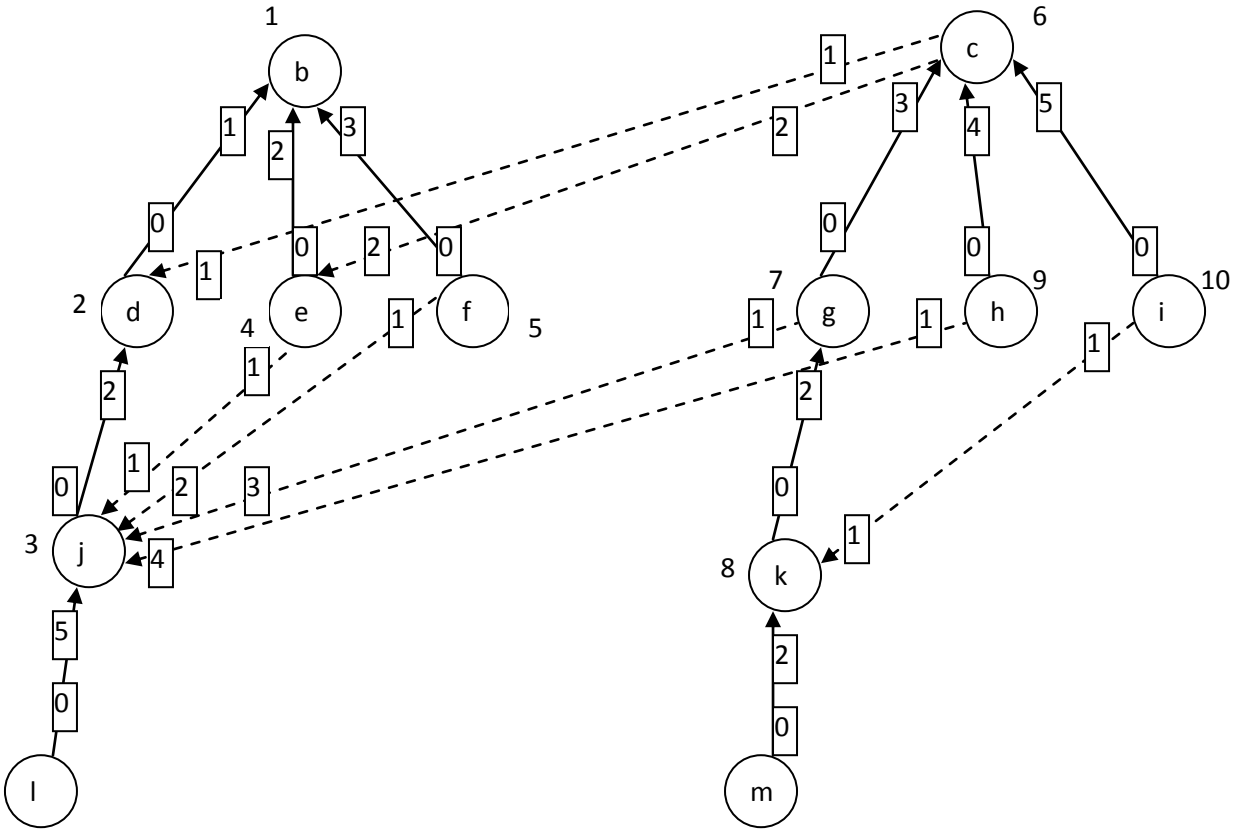


Fig 6 Activation of new switches 'l' and 'm' does not effect the graph G2 of Fig. 5 (new correct graph G3 after activation)

Table 5 Routing table for Fig 6 after activation of node 'l' and 'm' to Fig. 5

Switch	b	c	d	e	f	g	h	i	j	k	l	m
b	[X]	[1,2]	[1]	[2,1]	[3,1]	[1,2]	[1,2]	[1,2]	[1]	[1,2]	[1]	[1,2]
c	[1,2]	[X]	[1,2]	[2,1]	[2,1]	[3,2,1]	[4,2,1]	[5,2,1]	[1,2]	[3,2,1]	[1,2]	[3,2,1]
d	[0]	[1,2,0,]	[X]	[0,2]	[0,2]	[2,1,0]	[2,1,0]	[1,0,2]	[2]	[1,2,0]	[2]	[2,1,0]
e	[0,1]	[2,1,0]	[0,1]	[X]	[0,1]	[2,1,0]	[2,1,0]	[2,1,0]	[1,0]	[2,1,0]	[1,0]	[1,2,0]
f	[0,1]	[0,1]	[0,1]	[0,1]	[X]	[1,0]	[1,0]	[1,0]	[1,0]	[1,0]	[1,0]	[1,0]
g	[0,1]	[0,1]	[0,1]	[0,1]	[1,0]	[x]	[0,1]	[0,2,1]	[1,0]	[2,1,0]	[1,0]	[2]
h	[0,1]	[0,1]	[0,1]	[0,1]	[1,0]	[0,1]	[X]	[0,1]	[1,0]	[0,1]	[1.0]	[0,1]
i	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[X]	[1,0]	[1,0]	[1,0]	[1,0]
j	[0]	[0,1]	[0]	[1,0]	[2,0]	[3,0]	[4,0]	[3,1,0]	[X]	[3,1,0]	[5]	[3,1,0]
k	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[1,0]	[0]	[X]	[0]	[2]
l	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[X]	[0]
m	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[X]

cycle in G2, we have three options with respect to break node placement. Thus, it is always possible to construct a correct graph G2 which is also like G1. A correct graph G1 in Fig 4 gives a new correct graph G2 in Fig 5 after removal of root node 'a' with new root node 'b' OR correct graph G3 as shown in Fig 6 from graph G2 while addition of new nodes 'l' and node 'm'

Lemma 2

Assume that two up/down directed graphs G1 and G2 where G2 is also correct after removal of any node or link of correct graph G1

Proof

Assume that G2 is a subgraph of G1 in which each link that connects a break node in G1 with the corresponding break node in G2 is suppressed, and that G2 is a correct UP/DOWN graph. Then, according to lemma 1, we can guarantee that G2 is also a correct UP/DOWN graph. Thus it possible to define a fully connected deadlock-free routing function R2 over G2. R2 satisfies the routing-restrictions of both G1 and G2 since all the break nodes either have the same locations or have been removed.

Preorder Traversal Based UP/DOWN direction assignment to various link.

1) Computing a spanning tree.
2) Preorder traversal of spanning tree and assign labeling to nodes through horizontal traversal based distance of each node from root node.
3) Next assigning UP/DOWN directions to various links likewise previous strategy.
4) Finally, updation of routing tables

Switch Deactivation

The deactivation of switch including the root node produces a new root node. Switch deactivations imply that messages routed to remove components must be discarded. In this case, shorter reconfiguration time implies less discarded messages which is the characteristic of this improved PPR scheme.

Switch Activation

When a new switch is added, a direction must be assigned to the links connecting to it in such a way that the down direction goes toward the new switch and it should no produce cycles in the directed graph.

IV. Performance Evaluation

In this section, we evaluate the performance of the improved PPR algorithm proposed in section 3. Our dynamic reconfiguration scheme is compared to traditional PPR scheme. New mechanism requires very less computations time for updation of routing function from old one to new, when there is change in network topology due to addition or removal of switch nodes or links.

a) *Switch Model*

A switch consist of crossbar, a routing and configuration unit, and many full-duplex links. The routing and configuration unit provides the output channel for multiple packets to cross a switch. Table look-up routing is used. Each input channel has 2 set of buffers: user and control buffers. Control buffers handle control messages generated in each reconfiguration process when there is any change in network topology. We have assumed that one clock cycle is required to access the routing table and provide the output link for the message.

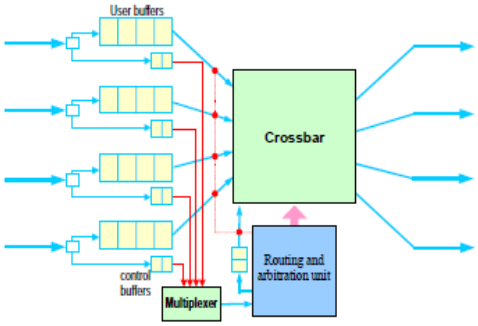


Fig.7 Switch Architecture

b) *Network Model*

The high-speed switched networks are consist of a set of switches interconnected through point-to-point links and hosts are connected to those switches through a network interface card in a irregular fashion. We have evaluated the performance of different sizes of networks. Several different networks with irregular topologies are considered in order to perform a detailed study. The irregular topologies have been randomly generated.

c) *Message Generation*

For each simulation run, we have considered that message generation rate is constant and the same for all the nodes. Once the network has reached a steady state, the flit generation rate is equal to the flit reception rate(traffic).We have evaluated the full range of traffic, from low load to saturation. On the other hand, we have considered the message destination is randomly chosen among all the nodes

d) *Simulation Results*

In this section, we show the simulation results for improved PPR scheme based on pre-order traversal of spanning tree. The simulation results are compared with existing PPR scheme. The simulation used for this work is developed with the IRFlexiSim Simulator[18]. First of all, we have discussed the way in which the path computation time is reduced by using improved PPR strategy in section 3. In order to increase the accuracy of the results, each experiment is repeated several times. The numbers of simulation runs for each topology are presented graphically. The number of data packets that are discarded during the topology change assimilation process gives an indication of the level of service a network can provide to applications. Fig. 8 compares the amount of packets that are discarded for PPR and improved PPR scheme. The results in Fig 8 shows that, for a switch removal the rate at which data packets are discarded is notably lower for new PPR scheme. The main reason is that, for improved PPR scheme minimum changes are required to update the routing tables. In case of switch additions, no packets

are discarded due to inactive ports because no old roots include the switch that was just added. There is no packet discarding when a switch addition occur because it does not destroy any of the existing paths. To conclude this evaluation, Figures as shown below illustrates the instantaneous behavior of both the old PPR and new PPR scheme. The improved PPR scheme reduces the number of packets discarded during reconfiguration and increases the performance by distributing the traffic over all links of the network. Thus, it reduces the latency of packets and increases throughput by optimize use of all links and lowering the congestion at root node as shown in Fig 9.

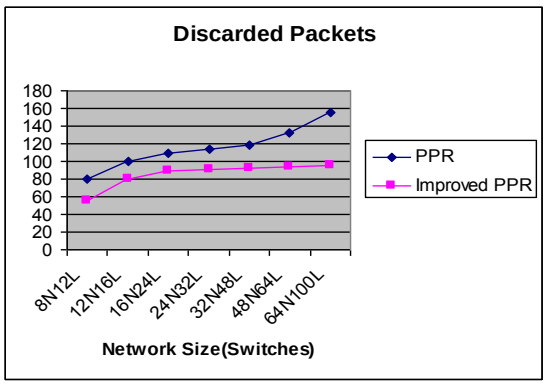


Fig. 8 Discarded Packets versus network size

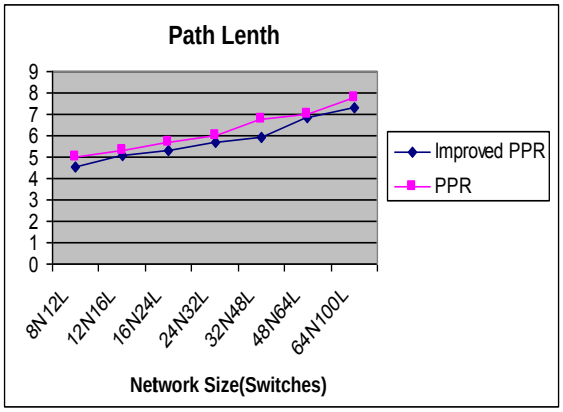


Fig. 9 Average path length versus network size

V. Conclusions and Future Work

Use In this paper, we have proposed and evaluated a simple improved reconfiguration strategy to compute the UP/DOWN routing tables. The improved methodology makes use of pre-order traversal of spanning tree in order to assign UP and DOWN directions to links is able to impose less routing restrictions to remove the cyclic channel dependencies in the network than those imposed when using the original methodology. Imposing less routing restrictions causes most messages to be routed through minimal paths and a better traffic balance, thus increasing channel utilization in the network .At a minor computational cost, the new routing function is designed to ensure that packets routed according to the old and the new routing functions can unrestrictedly coexist in the network, without the risk of forming deadlocks. Simulation results show that this significantly reduces the amount of packets that are discarded during the topology change assimilation. Moreover, our strategy does not required additional resources such as virtual channels, and it could easily be implemented in current commercial systems. As future work, we plan to extend the proposed methods in order to support other routing algorithms in addition to UP/DOWN.

References Références Referencias

1. Advanced Switching Interconnect Special Interest Group, Advanced Switching Core Architecture Specification Revision 1.0 ,December 2003, http://www.picmg.org
2. Boden, N.J., et al.: Myrinet: A gigabit per second LAN. IEEE Micro. ,February 1995
3. Casado, R., Bermúdez, A., Quiles, F.J., Sánchez, J.L., Duato, J.: A protocol for deadlockfree dynamic reconfiguration in high-speed local area networks. IEEE Transactions on Parallel and Distributed Systems 12(2) ,February 2001
4. Dally, W.J., Seitz, C.L.: Deadlock-free message routing in multiprocessor interconnection networks. IEEE Transactions on Computers 36(5) ,1987
5. InfiniBand Architecture Specification (1.2) (November 2002), http://www.infinibandta.com/
6. Lysne, O., Duato, J.: Fast dynamic reconfiguration in Irregular networks. In: Proc. Int.Conference on Parallel Processing ,August 2000
7. Lysne, O., Montañana, J.M., Pinkston, T.M., Duato, J., Skeie, T., Flich, J.: Simple deadlock-free dynamic network reconfiguration. In: Proc. of the International Conference on High Performance Computing,December 2004
8. Myrinet, Inc.: Guide to Myrinet-, Switches and Switch Networks ,2000 http://www.myri.com/
9. R.Horst, Tnet: A reliable System area network, IEEE Micro, Feb. 1995 ,pp 36-44
10. Pinkston, T.M., Pang, R., Duato, J.: Deadlock-free dynamic reconfiguration schemes for increased network dependability. IEEE Transactions on Parallel and Distributed Systems 14(6) ,June 2003
11. Robles -Gómez, A., García, E.M., Bermúdez, A., Casado, R., Quiles, F.J.: A Model for the Development of ASI Fabric Management

Protocols. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128. Springer, Heidelberg ,2006)

12. Robles-Gómez, A., Bermúdez, A., Casado, R., Quiles, F.J., Skeie, T., Duato, J.: A proposal for managing ASI fabrics. Journal of System Architecture (JSA) , 2008
13. Rodeheffer, T.L., Schroeder, M.D.: Automatic reconfiguration in Autonet. In: SRC Research Report 77 of the ACM Symposium on Operating Systems Principles ,October 1991
14. Sancho, J.C., Robles, A., Duato, J.: A new methodology to compute deadlock-free routing tables for irregular networks. In: Proc. the 4th Workshop on Communication, Architecture, and Applications for Network-based Parallel Computing ,January 2000
15. Schroeder, M.D., Birrell, A.D., Burrows, M., Murray, H., Needham, R.M., Rodeheffer, T.L., Satterthwate, E.H., Thacker, C.P.: Autonet: a high-speed, self-configuring local area network using point-to-point links. IEEE Journal on Selected Areas in Communications 9(8) ,October 1991
16. NetRec[] Avresky, Dimiter(2000) Dependable Network Computing(Kluwer Academic Publishers, Dordrecht), Chapter 10.
17. D.R. Avresky, Y.Varoglu and M.Marinov,"Dynamically Scaling System Area Networks", IEEE Proceedings of the 12th EuroMicro Conference On Parallel , Distributed and Network-Based Processing , , Feb 2004, pp.117 – 129.
18. http://www.ceng.usc.edu/smart/tools.html
19. Guillermo Ibanez ,Alberto Garcia –Martinez ,Juan A. Carral , Pedro A. Gonzalez ,Arturo Azcorra , Jose M. Arco(2010) "HURP/HURBA: Zero–configuration hierarchical Up/Down routing and bridging architecture for Ethernet backbones and campus networks" Elsevier Journal of Computer Networks (54) ,pp. 41-56

Global Journals Inc. (US) Guidelines Handbook 2011

www.GlobalJournals.org

FELLOWS

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.

AUXILIARY MEMBERSHIPS

ANNUAL MEMBER

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

PAPER PUBLICATION

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

Process of submission of Research Paper

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

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

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

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

(II) Choose corresponding Journal.

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

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

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

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

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be 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 .

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Informal Guidelines of Research Paper Writing

Key points to remember:

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

Final Points:

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

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

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

General style:

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

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

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

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

In every sections of your document

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

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

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

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

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

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

Title Page:

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

Abstract:

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

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

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

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

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

Approach:

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

Introduction:

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

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

Approach:

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

Procedures (Methods and Materials):

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

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

Materials:

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

Methods:

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

Approach:

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

What to keep away from

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

Results:

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

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

Content

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

What to stay away from

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

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

Approach

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

Figures and tables

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

Discussion:

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

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

Approach:

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

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

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

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

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

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

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

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

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

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

Index

U

W

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