Computer Science & Information Technology 27

David C. Wyld
Natarajan Meghanathan (Eds)

Computer Science & Information Technology

First International Conference on Computer Science & Information Technology (CoSIT 2014)
Bangalore, India, September 13 ~ 14 - 2014

AIRCC

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

ISSN: 2231 - 5403
ISBN: 978-1-921987-12-0
DOI : 10.5121/csit.2014.4901 - 10.5121/csit.2014.4924

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

First International Conference on Computer Science & Information Technology (CoSIT 2014) was held in Bangalore, India, during September 13~14, 2014. First International Conference on Data Mining (DMIN 2014), First International Conference on Signal and Image Processing (Sigl 2014), First International Conference on Cybernetics & Informatics (CYBI 2014), First International Conference on Networks, Mobile Communications & Telematics (NMCT 2014) and First International Conference on Artificial Intelligence and Applications (AIApp 2014) were collocated with the CoSIT-2014. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CoSIT-2014, DMIN-2014, Sigl-2014, CYBI-2014, NMCT-2014, AIApp-2014 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, CoSIT-2014, DMIN-2014, Sigl-2014, CYBI-2014, NMCT-2014, AIApp-2014 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CoSIT-2014, DMIN-2014, Sigl-2014, CYBI-2014, NMCT-2014, AIApp-2014.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld
Natarajan Meghanathan

Organization

Program Committee Members

Abd El-Aziz Ahmed	Cairo University, Egypt
Abdelouahab Moussaoui	Ferhat Abbas University - Sétif I, Algeria
Abdurrahman Celebi	Beder University, Albania
Adamu Murtala Zungeru	Federal University, Nigeria
Ah-Lian Kor	Leeds Metropolitan University, UK
Akhil Garg	Nanyang Technology University, Singapore
Alireza Sahab	Islamic Azad University, Iran
Amani K Samha	Queensland University of Technology, Australia
AmirReza	Islamic Azad University, Iran
Arash Habibi Lashkari	University Technology of Malaysia, Malaysia
Awais Azam	Middlese University, UK
Ayad Ghany Ismaeel	Hawler Polytechnic University , Iraq
Ayush Singhal	University of Minnesota, USA
Dac-Nhuong Le	Vietnam National University, Vietnam
Dananjayan P	Pondicherry Engineering College , India
Daniela Lopez De Luise	Universidad de Palermo, Argentina
Durgesh Samadhiya	Chung Hua University, Iran
Durjoy Majumder	West Bengal State University, India
Ederval Pablo Ferreira da Cruz	Federal Institute of Espirito Santo, Brazil
EL-Rabaie S	Menofia University, Egypt
Hazem Al-Najjar	Misrata University, Libya
Hossein Jadidoleslamy	Mangosuthu University of Technology, Iran
Hyung-Woo Lee	Hanshin University, South Korea
Ibrahim Alsonosi Nasir	Sebha University, Libya
Inukollu Narasimha	Texas Tech university , USA
Isa Maleki	Islamic Azad University, Iran
Jabbar MA	Aurora's Engineering College, India
Jagadeesh HS	A P S College of Engineering, India
Jalali M.H	Isfahan university of technology, Iran
Jamel Ghouraf	University of SBA, Algeria
Jan Zizka	Mendel University , Czech Republic
Joberto S B Martins	Salvador University, Brazil
John Tengviel	Sunyani Polytechnic, Ghana
Khanbabaei	Islamic Azad University, Iran
Kwan Hee Han	Gyeongsang National University, South Korea
Lallie Harjinder	University of Warwick, UK
Luisa B. Aquino	University of Saint Louis, Philippines
Marco Folli	University of Pavia , Italy
Mohammad Masdari	Islamic Azad University, Iran
Mohammed Arif Amin	Higher Colleges of Technology , Australia
Narasimha I.V	Texas Tech University, USA
Natarajan Meghanathan	Jackson State University , USA
Noureddine Bouhmala	Buskerud and Vestfold University, Norway

Peyman Mohammadi	Islamic Azad University, Iran
Phuc V. Nguyen	EPMI , France
Prasad T. V	Visvodaya Technical Academy, India
Rahali Bouchra	University of Tlemcen, Algeria
Saad M. Darwish	Alexandria University, Egypt
Sarmistha Neogy	Jadavpur University , India
Sary Awad	Ecole des Mines de Nantes, France
Seifedine Kadry	American University of the Middle East, Kuwait
Seyyed AmirReza Abedini	Islamic Azad University, Iran
Seyyed Reza Khaze	Islamic Azad University, Iran
Sundarapandian Vaidyanathan	Vel Tech University , India
Suvineetha Herath	University of Bahrain, Kingdom of Bahrain
Tad Gonsalves	Sophia University , Japan
Tinatin Mshvidobadze	Gori University, Georgia
William Simpson	Institute for Defense Analyses, USA
Yeongdeok Kim	Yonsei University, Korea

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

First International Conference on Computer Science & Information Technology (CoSIT 2014)

First International Conference on Data Mining (DMIN 2014)

First International Conference on Signal and Image Processing (Sigl 2014)

First International Conference on Cybernetics & Informatics (CYBI 2014)

First International Conference on Networks, Mobile Communications & Telematics (NMCT 2014)

First International Conference on Artificial Intelligence and Applications (AIApp 2014)

SPECTRAL RELEVANCE CODING IN MINING FOR GENOMIC APPLICATION

S.J.Saritha[1] and Prof.P.Govindarajulu[2]

[1]Dept of CSE, JNTUA CE Pulivendula,A.P , INDIA
sarithajntucep@gmail.com
[2]Dept of CSE, S.V.University, Tirupathi A.P., INDIA,
pgovindarajulu@yahoo.com

ABSTRACT

Most current gene detection systems are Bio-informatics based methods. Despite the number of Bio-informatics based gene detection algorithms applied to CEGMA (Core Eukaryotic Genes Mapping Approach) dataset, none of them have introduced a pre-model to increase the accuracy and time reduction in the different CEGMA datasets. This method enables us to significantly reduce the time consumption for gene detection and increases the accuracy in the different datasets without loss of Information. This method is based on feature based Principal Component Analysis (FPCA). It works by projecting data elements onto a feature space, which is actually a vector space that spans the significant variations among known data elements.

KEYWORDS

Gene detection system, PCA, KPCA, Spectral simulation

1. INTRODUCTION

Communication networks make physical distances worthless. People can communicate with each other through the networks without any restriction of the real distance. While we treasure the ease of being connected, it is also recognized that a gene users from one place can cause severe damages to wide areas. Generally a gene is defined as "any set of actions that attempt to compromise the integrity, confidentiality or availability of information resources." The identification of such a set of malicious actions is called gene detection problem. The Gene detection systems are an integral package in any well configured and managed computer system or network. Generally Gene detection systems may be some software or hardware systems that monitor the different events occurring in the actual system and analyzing them for effective detection.

There are two major approaches in gene detection: anomaly detection and misuse detection. Misuse detection consists of first recording and representing the specific patterns of genes, then monitoring current applications for such patterns, and reporting the matches. There are several developed models in misuse gene detection [1] [2]. They differ in representation as well as the matching algorithms employed to detect such threat patterns. Anomaly detection, on the other hand, consists of building models from normal data and then detects variations from the normal model in the observed data. The main advantage with anomaly gene algorithms is that they can

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 01–07, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4901

detect new forms of genes, because these new genes will probably deviate from the normal original behavior of genes [3].

There are many Gene detection systems developed for gene detection. But most of them apply an algorithm directly [4, 5, 6] on the rough data obtained from traffic or other local or remote applications which increases the consumption time. The CEGMA gene detection datasets [7] are an example for these algorithms. To overcome the draw back of high time consumption, a method was proposed for gene detection based on the principal component analysis (PCA) [8]. This method Extracts the main components (repetitive components) of the incoming dataset and performs the gene detection only for those components. However this method reduces the time consumption but reduces the accuracy. To overcome this drawback another method is proposed named as advanced PCA. This method accomplishes with the clusters of incoming dataset based upon their header information. Though this method increases the accuracy and reduces the time consumption but there is possibility to alter the incoming bio informatics at switching stages. Thus it can be considered as valid. To overcome this drawback there should be another parameter to analyze the incoming informatics. This paper proposes a method to overcome the drawback of previous method by introducing a new parameter called spectral simulation. This method performs the calculation of spectral nature of incoming gene data set if the header of the incoming data packet is not matched. The rest of this paper is organized as follows;

Section II gives the detailed description of PCA on the gene data set. Section III gives the cluster formation of incoming gene data based on the specific features. Proposed spectral simulation method is discussed in section IV. The results obtained are represented in section V and finally section VI concludes the paper.

2. PRINCIPAL COMPONENT ANALYSIS (PCA)

This section gives the complete illustration about the principal component analysis and also tells how to extract the important (repetitive) features of the complete incoming gene dataset. It is often used to reduce the dimension of dataset for easy exploration. It is concerned with explaining the variance-covariance structure of a set of variables through a few new variables which are functions of the original variables. Principal components are particular linear combinations of the p random variables. Let X_1, X_2, …, X_p be the P random variables representing p gene datasets with three important properties: (1) the principal components are uncorrelated, (2) the first principal component has the highest variance, the second principal component has the second highest variance, and so on, and (3) the total variation in all the principal components combined is equal to the total variation in the original variables X_1, X_2, …, X_p. They are easily obtained from an Eigen analysis of the covariance matrix or the correlation matrix of X_1, X_2, …, X_p [9].

Principal components from the covariance matrix and the correlation matrix are usually not the same. In addition, they are not simple functions of the others. When some variables are in a much bigger magnitude than others, they will receive heavy weights in the leading principal components. For this reason, if the variables are measured on scales with widely different ranges or if the units of measurement are not commensurate, it is better to perform PCA on the correlation matrix.

Let **R** be a p x p sample correlation matrix computed from n observations on each of p gene datasets X_1, X_2, …, X_p. If (λ_1, $\mathbf{e}_1$), (λ_2, $\mathbf{e}_2$), …, (λ_p, $\mathbf{e}_p$) are the p eigen value-eigenvector pairs of **R**, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, then the i^{th} sample principal component of an observation vector x=(x1,x2,…..xp) is

$$y_i = \mathbf{e}_i'\mathbf{z} = e_{i1}z_1 + e_{i2}z_2 + \ldots + e_{ip}z_p, \quad i = 1,2,\ldots,p$$ Where

$\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{ip})'$ is the i^{th} eigenvector

And

$\mathbf{z} = (z_1, z_2, \ldots, z_p)'$ is the vector of standardized observations defined as

$$z_k = \frac{x_k - \overline{x}_k}{\sqrt{s_{kk}}}, \quad k = 1,2,\ldots,p$$

Where k_x and s_{kk} are the sample mean and the sample variance of the variable X_k.

The i^{th} principal component has sample variance λ_i and the sample covariance of any pair of principal components is 0. In addition, the total sample variance in all the principal components is the total sample variance in all standardized variables $Z_1, Z_2, \ldots, Z_p$, i.e., $\lambda_1 + \lambda_2 + \ldots + \lambda_p = p$

This means that all of the variation in the original dataset is accounted by the principal components. But this method allows only repetitive components to classify with the incoming data set at testing. This is very effective in reducing the computation time by decreasing the total size where as the main draw back of this method, it is not able to give accuracy because when there is data set testing which is not a repetitive one. To overcome this problem a cluster based PCA is proposed and also discussed briefly in next section.

3. K-PCA

This section provides the information about the PCA based on kernel (clusters) features. Like in PCA, the overall idea is to perform a transformation that will maximize the variance of the captured variables while minimizing the overall covariance between those variables. Using the kernel trick, the covariance matrix is substituted by the Kernel matrix and the analysis is carried analogously in feature space. An Eigen value decomposition is performed and the eigenvectors are sorted in ascending order of Eigen values, so those vectors may form a basis in feature space that explain most of the variance in the data on its first dimensions.

However, because the principal components are in feature space, we will not be directly performing dimensionality reduction. Suppose that the number of observations **m** exceeds the input dimensionality **n**. In linear PCA, we can find at most **n** nonzero Eigen values. On the other hand, using Kernel PCA we can find up to **m** nonzero Eigen values because we will be operating on an ***mxm*** kernel matrix [10]. When the external features of all variables are matched with the features of variables present in database the gene is said to be detected, otherwise the variables are allowed for further process. Though this method increases the accuracy and reduces the time consumption but there is possibility to alter the incoming bio informatics of the gene dataset at various switching stages. Thus it can be considered as valid gene. To overcome this drawback there should be another parameter to analyze the incoming informatics. The next section gives the information about the spectral properties of incoming gene dataset which are allowed to further process.

4. PROPOSED METHOD

This method overcomes the above mentioned problem by extracting the spectral features of the incoming gene dataset. This method allows the comparison of spectral features of incoming dataset along with the normal features. In this method the internal features of the incoming gene dataset are also going to be compared with the features of dataset in the database. Then only they are going to allow for further process. Before this the complete spectral features of the gene dataset are have to be evaluated. For this purpose the complete incoming dataset is going to be represented in Binary format (1's and 0's).After this each and every gene is represented with a bit vector. This paper assumes the spectral characteristics of a gene data set as,

1. No of switching states out of all bits. I.e. how much number of times the bits changed their state out of all bits.
2. The symmetry property.
3. The transition time taken from one bit to next bit.

This spectral property plays a vital role in this paper. Based upon these spectral properties the incoming dataset is going to be tested and allowed for further process The data set present in the personal computer is divided into clusters based upon their headers as shown below.

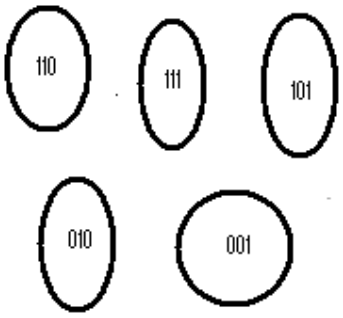
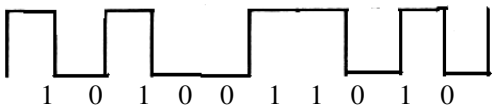


Figure1: clusters of data base

The incoming data set is compared with these clusters. If the header of a incoming data set is matched with any one of the clusters header it is detected as valid gene. This is processed out in previous approach.

In this approach first the present dataset is divided into clusters and also their spectral characteristics are calculated as follows:

Let a gene is represented with the bit vector shown below



The total number of bits=10
The total number switching states=7
The total switching ratio =7/10=.7

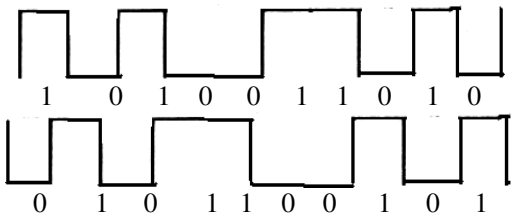


Figure2: spectral properties of dataset

Like this the total switching ratio is calculated for each and every data packet and kept in a cluster. When the incoming data set is said to matched with the header information of any cluster the spectral properties of that incoming dataset is also going to compared with spectral properties of the dataset. If they are matched then the incoming data set is said to be gene otherwise it is allowed for further process. The data is going to be switched by many steps during transmission. So there is possibility to change the header information intentionally and also non-intentionally. Non-intentionally means automatic change of header during transmission. There may be a possibility to change the header information by hackers also. This is referred to as intentional change. So the comparison of spectral properties of incoming dataset increases the accuracy as well as reduces the time consumption. The results discussed below gives the graphical information about this proposed method.

5. RESULTS

This section gives the illustration about the performance evaluation of the proposed method.

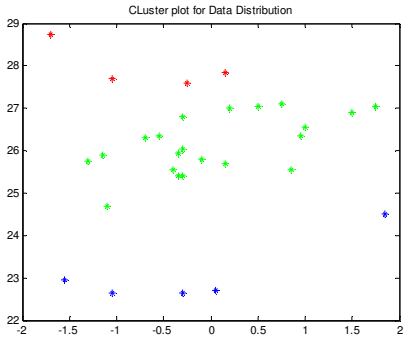


Figure 3: data scattering plot for relevance gene sequence

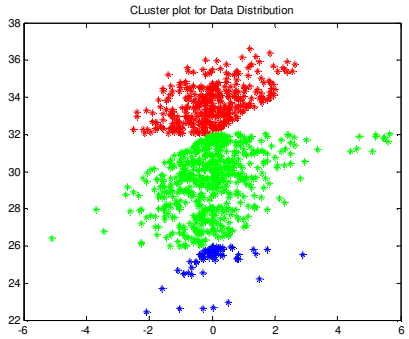


Figure 4: relevance clustering of genomic information in k class

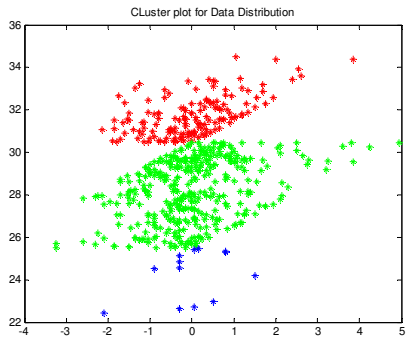


Figure5: relevance clustering of genomic data set for k-class after spectral mapping

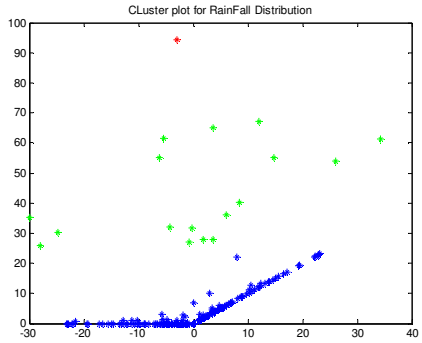


Figure6: cluster relevancy for a single class observation in 3–set data

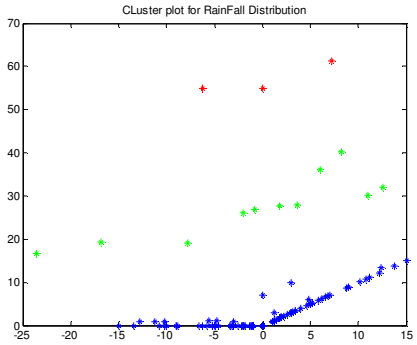


Figure 7: relevance plot of 1-class among 3-class Clustered data at spectral mapping

6. CONCLUSION

To improve the operation of data mining in this paper a spectral based doing approach is proposed. The proposed approach observes the variation in sequence pattern is developed and in similarity to a spectral correlation is observed. Pattern having sequence of similar spectral information is defined in bit pattern and a similar code is applied for representation to the existing coding pattern. For the test of such approach extended format of PCA called Kernel- PCA (K-PCA) is used. From the obtained observations it is observed that a improvement in processing efficiency with respect to Process time and recall efficiency is observed.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] K. Ilgun. Ustat, a real time gene detection system for UNIX. In IEEE Symposium on Security and Privacy, pages 16–28, Oakland, CA, May 1993.

[2] S. Kumar and E. Spafford. A pattern matching model for misuse gene detection. In Proceedings of the 17th National Computer security Conference, pages 11–21, 1994.

[3] D. Denning. An Gene Detection Model. IEEE Transactions on Software Engineering, 13(2):222–232, 1987.

[4] R. Agrawal and M. V. Joshi. PNrule: A New Framework for Learning Classifier Models in Data Mining A Case-Study in Network Gene detection. Technical Report RC-21719, IBM Research Division, 2000.
[5] I. Levin. CEGMA-99 Classifier Learning Contest LLSoft's Results Overview. SIGCEGMA Explorations. ACM SIGCEGMA, 1:67–71, 2000.
[6] B. Pfahringer. Winning the CEGMA Classification Cup: Bagged Boosting. SIGCEGMA Explorations.ACM SIGCEGMA, 1:65–66, 2000.
[7] CEGMA Cup 99 Gene Detection Datasets. Available at: http://CEGMA.ics.uci.edu/databases/CEGMAcup99/CEGMAcup99.html, 1999.
[8] I. T. Joliffe. Principal Component Analysis. Springer Verlag, New York, NY, third edition, July 2002.
[9] I.T.Jolliffe,"principal component analysis", Ed.,springer-verlag, NY, 2002
[10] FASEL, Ian. Scholkopf, Smola and Muller: KernelPCA. Available in: http://cseweb.ucsd.edu/classes/fa01/cse291/kernelPCA_article.pdf
[11] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.
[12] Gizem, Aksahya & Ayese, Ozcan (2009) Coomunications & Networks, Network Books, ABC Publishers.

AUTHORS

Smt. S.Jessica Saritha is currently working as an Assistant Professor in Department of CSE JNTUA College of Engineering, Pulivendula,Andhra Pradesh India Her research interests are Data mining and Distributed databases

Prof. P. Govindarajulu is a rertired professor in department od Computer Science and Engineering Sri. Venkateswara University Tirupathi, Andhra Pradesh . HE worked at various portfolios in the university . His research interests of the Databases and Data mining

INTENTIONAL BLANK

CASE BASED STUDY TO ANALYZE THE APPLICABILITY OF LINEAR & NON-LINEAR MODELS

Anubhav Gupta[1], Gaurav Singh Thakur[2] , Ankur Bhardwaj[3] and Biju R Mohan[4]

Department of Information Technology,
National Institute of Technology Karnataka, Surathkal
[1]Anubhav992@gmail.com
[2]Sai007gaurav@gmail.com
[3]bhardwajankur3@gmail.com
[4]biju@nitk.ac.in

ABSTRACT

This paper uses a case based study – "product sales estimation" on real-time data to understand the applicability of linear and non-linear models. We use a systematic approach to address the given problem statement of sales estimation for a given product by applying both linear and non-linear techniques on a data set of selected features from the original data set. Feature selection is a process that reduces the dimensionality of the data set by eliminating those features which contribute minimal to the prediction of the dependent variable. The next step is training the model which is done using two techniques from linear & non-linear domains, one of the best ones in their respective areas. Data Re-modeling has then been done to extract new features from the data set by changing the structure of the dataset & the performance of the models is checked again. Data Remodeling often plays a crucial role in boosting classifier accuracies by changing the properties of the dataset. We then try to analyze the reasons due to which one model proves to be better than the other & hence try and develop an understanding about the applicability of linear & non-linear models. The target mentioned above being our primary goal, we also aim to find the classifier with the best possible accuracy for product sales estimation in the given scenario.

KEYWORDS

Machine Learning, Prediction, Linear and Non-linear models, Linear Regression, Random Forest, Dimensionality Reduction, Feature Selection, *Homoscedasticity.*

1. INTRODUCTION

Machine learning is a branch of artificial intelligence. It concerns the construction and study of systems that can learn from data. For example, a machine learning algorithm can be used to classify people by gender, by data such as height, Body Mass Index, Favorite color etc. There are various types of Machine Learning Algorithm; two of the main types are Supervised Learning

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 09–19, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4902

(where the desired output is Known), and Unsupervised Learning (where the desired output is not known). In this paper, we discuss the techniques which belong to Supervised learning[3].

Predicting the future sales of a new product in the market has intrigued many scholars and industry leaders as a difficult and challenging problem. It involves customer sciences and helps the company by analyzing data and applying insights from a large number of customers across the globe to predict the sales in the upcoming time in near future. The success or failure of a new product launch is often evident within the first few weeks of sales. Therefore, it is possible to forecast the sales of the product in the near future by analyzing its sales in the first few weeks. We propose to predict the success or failure of each of product launches 26 weeks after the launch, by estimating their sales in the 26^{th} week based only on information up to the 13th week after launch. We intend to do so by combining data analysis with machine learning techniques and use the results for forecasting.

We have used the divided the work into following phases:

i) Dimensionality reduction (Feature selection)

ii) Application of Linear & Non-Linear Learning Models

iii) Data Re-modeling

iv) Re-application of learning models.

v) Evaluation of the performance of the learning models through comparative study & Normality tests.

vi) Boosting the accuracy of the model that better suits the problem based on their evaluation.

To create a forecasting system for this problem statement we gathered 26 weeks information for nearly 2000 Products belonging to 198 categories to train our model. Various attributes such as units_sold_that_week, Stores_selling_in_the_nth_week, Cumulative units sold to a number of different customer groups etc are used as independent variables to train & predict the dependent variable- "Sales_in_the_nth_week". However our task here is only to predict their sales in the 26^{th} week.

In Section 2, we discuss about the methodology and work done in each of the phases, followed by the results & discussion in Section 3. Finally we draw a conclusion in Section 4 along with its applications, followed by the references.

2. METHODOLOGY AND WORK DONE

A basic block diagram to explain the entire process of Machine Learning is given below.

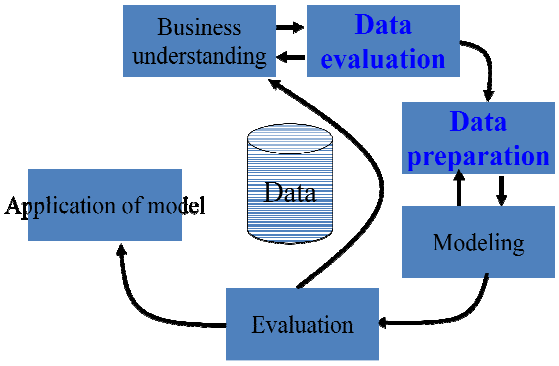


Figure 1 : Machine Learning Life Cycle

2.1 Feature selection

We used Greedy Stepwise[2] mechanism for feature selection[2]. The process of feature selection gives us a list of important features from the original feature set. Here stores_selling_in_the_nth_week and weeks_since_launch have been the two most important features with maximum sales predicting power in the original data set.

The results from this procedure can be backed up using the scatter plots. The scatter plots are used for the feature "Total Units sold in nth week" plotted against other features. Their variations are then studied and can be used as a reference to justify the results from feature selection. Those scatter plots with random nature can be easily identified and discarded.

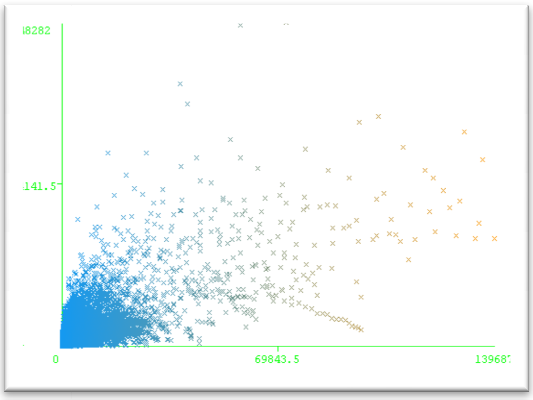


Figure 2: Cumulative Units sold to very price sensitive customers vs. Total units Sold

The above scatter plot is between:

y-axis	Total Units sold
x-axis	Units sold to Very Price sensitive customers

We can clearly see, the scatter plot between the two features does not show any trend as it is completely random in nature. A similar scatter plot was seen for most of the features, except for those obtained from feature selection. For the ones obtained from the feature selection process these scatter plots showed some relation between them which confirms the fact that they are the necessary features for regression.

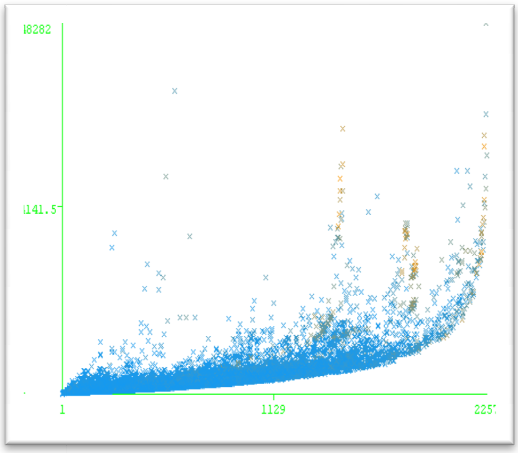


Figure 3 : Total Units Sold Vs Stores Selling

y-axis	Total Units sold
x-axis	Stores Selling

Hence, this allows us to reduce the number of features that must be used to train our model.

2.2 Linear Model

We used Multiple Linear Regression[1] for our linear learning model. The equation for Multiple Linear Regression is given below.

$$\mathbf{F_\Theta(X) = \Theta + \sum \Theta_i X} \qquad \text{(Eq.. 1)}$$

Where, X is the set of input vector with coefficients/weights Θ_i and constant value of Θ called the bias. $F_\Theta(X)$ is the approximated Linear function to be used for regression.

This model needs to be optimized by minimizing the Mean Square Error produced by the model. The cost function in this case is:

$$\mathbf{J(\Theta_0, \Theta_1)=(1/2m) \sum (F_\Theta (x_i)- y_i)^2} \qquad \text{(Eq.. 2)}$$

Where, $F_\Theta (x_i)$ is the predicted value, y_i is the actual value, and 'm' is the number of tuples used for training. This is the cost function which has been optimized using Gradient Descent Algorithm [4].

We have applied this linear learning model on the data set of selected features. The results obtained have been mentioned in the next section.

2.3 Non-Linear Model

We use Random Forest, a bagging based ensemble learning technique for non-linear training. A Random Forest[9] consists of a collection or ensemble of base decision tree predictors/classifiers, each capable of producing a response when presented with a set of predictor input values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable. Alternatively, for regression problems, the tree response is an estimate of the dependent variable given the predictors, taking the average from each tree as the net output of the model.

Each tree is grown as follows[3]:

1. Firstly, create n_tree bootstrap samples allowing selection with replacement, where each sample will be used to create a tree.

2. If there are M input variables, a number m<M is specified such that at every node, m variables are selected at random out of the M and the best splitting attribute off these m is selected. The value of m is kept constant during the process.

3. Each tree is grown completely without pruning.

This technique was implemented in R tool, with the parameter values as n_trees = 500 and m (variables tried as each split) = sqrt(Number of features).

The accuracy of Random forests is calculated from the out-of-bag MSE which provides an unbiased result and eliminates the need for cross-validation.

$$\mathbf{MSE = (1/n) \sum (y_i - F_{\Theta\_oob})^2} \qquad \text{(Eq.. 3)}$$

2.4 Data Remodeling

Data Remodeling is a phase that requires some domain specific knowledge and use of problem specific information to restructure the data. Another approach could be the Brute Force technique which is not a good practice. We have made use of certain basic assumptions related to the market activities to make changes in this stage. We progressively make changes to the data set and analyze their results with the aim to improve them further.

2.4.1 Stage 1

The Data set provided for the problem statement originally had the following structure.

- Independent Variables – product id, product category, weeks since launch, stores selling in that week and various sales data to categorical customers.
- Dependent Variables – The total sales in the nth week.

The current approach is basically dividing the dataset based on the “product_category” and training the model for each one of them separately. This goes by the intuition that the market

sales patterns and demands-supply varies differently for different categories. And hence we regress separately for each category and use that model to predict the sales of a product for the 26^{th} week.

After some study, we had already identified in the initial phase that, for a particular category of product, only the weeks since launch and number of stores selling have a major effect in predicting the total sales for that week. But this model was not exactly suitable as:

- Firstly, the sales in the 26^{th} week apart from stores selling are also dependent on the sales in the previous weeks which were not being considered in the previous data model.

- Secondly, since in the test cases, the data provided is only for 13 weeks, the training must also not include any consumer specific sales data from beyond 13 weeks.

- The independent variable to be predicted must be the "Total Sales in the 26^{th} week" and not the "Total Sales in the nth week".

Hence, we have modified the data set such that we use the sales in every week upto 13 weeks along with the stores selling in week 26 as a feature set to estimate the sales in week 26. This way we can also measure the predictive power of sales in each week and how do they affect the sales in the later stages. This has been analyzed using feature selection on the new set and also through performing an Autocorrelation analysis on the 'sales_in_nth_week' to find the correlation between the sales series with itself for a given lag. The acf value for lag 1 was 0.8 and for lag 2 was 0.6 showing that with so much persistence, there is a lot of predictive power in the Total sales in a given week that can help us predict the sales for atleast two more weeks. Finally, the new structure of the dataset for this problem statement is as follows:

- Independent Variables – Sales in week1, Sales in week2, Sales in week3… Sales in week13, stores_selling_in_the_13th_week

- Dependent variable - Total Sales in the 26th week.

This dataset was then subjected to both Linear & Non-linear learning models. The one which performs better would then be used to train on the next phase of Data Remodeling.

2.4.2 Stage 2

We needed to further modify the data structure to improve results and also find a method by which we could regress the data set for all the categories together. This meant trying to find a model that allowed us to train a single model that could work on all the categories together. To do this, we used the following strategy:

1. Let 'usn' represent Units_sold_in_week_n and 'ssn' represent Stores_selling_in_week_n.

2. Now, as we had previously obtained the hypothesis from Linear Regression (Eq.. 1), in the form of:

$$\mathbf{usn = (k) \times ssn} \qquad \text{(Eq.. 4)}$$

where k is the co-efficient of ssn. Note that k is the only factor which would vary from category to category.

3. Therefore, from Eq.. 4, we get
    1. **us26 = (k) x ss26** (Eq.. 5)
    2. **us13 = (k) x ss13** (Eq.. 6)
    3. **=> us26/ us13 = ss26/ ss13** (Eq.. 7)
    4. **=> us26= ss26/ ss13 * us13** (Eq.. 8)

4. In this way, we remove the need for finding the need of the Coefficient of **ssn** for each category and simply keep an additional attribute '**ss26/ ss13 * us13'.**

5. Also, instead of keeping only the "Stores_selling" in week 26, we decided to keep "Stores_selling" from week 14 to 26 to further incorporate the trend (if any) of the Stores_selling against Units_sold_in_week_26. This was the only useful feature provided beyond 13th week. The number of stores could help us identify the trends in the sales of the product hence further improving the accuracy of our predictions in the 26th week.

Hence the structure of our new dataset was as follows:

1. For each of weeks 1 to 13, the ratio of stores in week 26 to stores in week 13, multiplied by the sales in that week.
2. The raw sales in weeks 1 through 13
3. The number of stores in weeks 14 through 26.

The results obtained from these changes are explained later in the next section.

2.5 Understanding the Applicability of Models

Any Linear model can only be applied on a given dataset assuming that it encompasses the following properties, else it performs poorly.

1. **Linearity** of the relationship between dependent and independent variables.
2. **Independence** of the errors (no serial correlation).
3. **Homoscedasticity**[12] means that the residuals are not related to the variable plotted on X-axis.
4. **Normality** of the error distribution.

In this case we test these properties to understand and justify the performance of Linear Models against a Non-Linear Models in this domain. These tests are conducted by:

1. Linear relationship among the features is a domain based question. For example does the "sales to price sensitive customer" affect its "stores selling in nth week". Such errors can be fixed only by applying transformations that take into account the interactions between the features.

2. Independence of errors is tested by plotting the Autocorrelation graph for the residuals. Serial correlation in the residuals implies scope for improvement and extreme serial correlation is a symptom of a bad model.

3. If the variance of the errors increases with time, confidence intervals for out of-sample predictions tend to be unrealistically narrow. To test this we look at plots of *residuals versus time* and *residuals versus predicted value*, and look for residuals that increase (i.e., more spread-out) either as a function of time or the predicted value.

4. The best test for normally distributed errors is a *normal probability plot* of the residuals. This is a plot of the fractiles of error distribution versus the fractiles of a normal distribution having the same mean and variance. If the distribution is normal, the points on this plot fall close to the diagonal line.

The results obtained in these tests are given in the next section.

3. RESULTS AND DISCUSSION

3.1 Feature selection

The list of features obtained from Greedy Stepwise feature selection[2] showed that "Stores Selling in nth week" and "weeks since launch" were the most important features contributing to the prediction of sales. The variance of these features with the dependent variable, together is greater than 0.94 showing that they contribute the maximum to the prediction of sales.

3.2 Application of Linear Regression And Random Forest

Linear Regression Considering the top 6 features obtained from feature selection procedure based on their variances:

Correlation Coefficient	0.9339
Mean Absolute Error	28.37
RMSE	69.9397

Random Forests considering all the features:

OOB-RMSE	46.26

As we currently see, the non-linear model is working better than the linear model. This may lead to a jumpy conclusion that non-linear model is probably better in this scenario. Moreover the accuracy of the classifiers is also not great due to the high RMSE values of both the models.

3.3 Application of Learning Models after Data Re-modeling Phase-1

Linear Regression Results:

Correlation Coefficient	0.9972
Mean Absolute Error	0.4589
RMSE	0.9012

Random Forest Results:

OOB-RMSE	7.19

As we see here, the performance of both the models have improved drastically, however, we find that the linear model outperforms random forest. This finding compelled us to inquire about the properties of the dataset that satisfied the assumptions of the linear model. We found that:

i) The Franke's Anscombe[11] experiment to test the normality of data distribution came out inconclusive leading us to use the Normal Q-Q plot[13].

ii) The Normal Q-Q plot in R[14] concluded that the dataset follows the normal distribution.

iii) The residuals also follow the normal distribution curve under the Normal Q-Q plot just like the actual data conforming the second assumption of linearity.

iv) We check the Homoscedasticity[12] property by plotting the residuals against fitted values. The graph was completely random in nature.

v) Lastly, the linear relationship between features is a domain specific question. The data collected mostly contains the sales data from local stores, from local manufactures of items of daily consumption types like – bread, milk_packets, airbags, etc. Since these types of products belong to a class of items where the stochastic component is negligible, it makes it easy for us to assume that the linear model can be easily applied to this problem. This is the reason why linear model is working better compared to the non-linear model due to negligible interaction of the features.

3.4 Application of Linear Models after Data Re-modeling Phase-2

Linear Regression Results considering all the new features:

Correlation Coefficient	0.9994
Mean Absolute Error	0.3306
RMSE	0.4365

Linear Regression Results considering only top 6 new features after applying feature selection on the new dataset:

Correlation Coefficient	0.99983
Mean Absolute Error	0.408
RMSE	0.7021

As we see, the results have improved further, with the accuracy of the classifier going up from RMSE value of approximately 65 to 0.43. With the final model, we were able to predict the Total

sales of any given product in the test set with an error < 1 unit for any category, our best RMSE achieved being 0.43.

4. CONCLUSION

The primary target in machine learning is to produce the best learning models which can provide accurate results that assist in decision making, forecasting, etc. This brings us to the essential question of finding the best suitable model that can be applied to any given problem statement. We have performed a case based study here to understand on how to decide whether a linear or a non-linear model is best suited for a given application.

We initially follow a basic approach by adopting two leading classifiers from each domain and evaluate their performances. We then try to boost the accuracies of both the learning models using data re-structuring. The results obtained from this process help us derive an important empirical proof that the accuracy of a classifier not just depends on its algorithm. There is no such certainty that a more complex algorithm will perform better than a simple one. As we see in this case, Random Forests, which belong to the class of ensemble classifiers bagging based is known to perform well and produce high accuracies. However, here the simple Multiple Linear Regression model outperforms the previous one. The accuracy of the model largely depends on the problem domain where it is being applied and the data set, as the domain decides the properties that the data set would inherit and this greatly determines the applicability of any machine learning technique. Hence holding a prejudice for/against any algorithm may not provide optimal results in machine learning.

To further this observation, the use of various other algorithms such as Artificial Neural Networks (Which is known to give great results in case both Linear and Non-linear Machine learning problems), as well as Support Vector Machines (Which are known to give very high accuracies) are suggested.

The framework developed here has been tested on real-time data and has provided accurate results. This framework can be used for the forecasting of daily use products, items of everyday consumption, etc. from local manufacturers, as it follows the assumption that the features have minimum interaction with each other. Branded products from big manufacturers include many more market variables, like the effect of political and economic factors, business policies, government policies, etc. which increase the stochastic factor in the product sales & also increase the interaction among the independent features. This feature interaction is very minimal for local products. Extending this framework to the “branded” scenario will require significant changes. However, the current model is well suited to small scale local products and can be easily used with minimal modifications, for accurate predictions.

REFERENCES

[1] Jacky Baltes, Machine Learning Linear Regression , University of Manitoba , Canada
[2] Isabelle Guyon and Andrė Elisseeff ,An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) 1157-1182
[3] Jiawei Han and Micheline Kamber, Data Mining –Conepts and Techniques , Second Edition Page-79-80
[4] Kris Hauser , Document B553, Gradient Descent, January 24,2012

[5] Luis Carlos Molina, Lluís Belanche, Àngela Nebot ,Feature Selection Algorithms: A Survey and Experimental Evaluation, University Politècnica de Catalunya
[6] Quang Nhat Nyugen, Machine Learning Algorithms and applications, University of Bozen-Bolzano, 2008-2009
[7] Jan Ivar Larsen, Predicting Stock Prices Using Technical Analysis and Machine Learning, Norwegian University of Science and Technology
[8] Classification And Regression By Random Forest by Andy Liaw And Matthew Wiener, Vol 2/3, December 2002
[9] Lecture 22,Classification And Regression Trees, 36-350, http://www.stat.cmu.edu/, November 2009
[10] [Online] Anscombe's Quartet- http://en.wikipedia.org/wiki/Anscombe's_quartet
[11] [Online] Homoscedasticity- http://en.wikipedia.org/wiki/Homoscedasticity
[12] [Online] Normal Q-Q Plot- http://en.wikipedia.org/wiki/Q–Q_plot
[13] [Online] Quick R-http://www.statmethods.net/advgraphs/probability.html

AUTHORS

Biju R Mohan
He is an Assistant Professor at National Institute of Technology Karnataka, Surathkal in the department of Information Technology. His areas of interest are Software Aging, Virtualization, Software Engineering, Software Architecture, Software Patterns, Requirements Engineering.

Gaurav Singh Thakur
Gaurav Singh Thakur completed his B.Tech in Information Technology from National Institute of Technology Surathkal and is currently working as a Software Engineer at Cisco Systems, Inc. Bangalore. His technical areas of interest include Machine learning, Networking & Security and Algorithms.

Anubhav Gupta
Anubhav completed his Bachelors , in National Institute of Technology Karnataka, Surathkal in the Field of Information Technology (IT). His areas of interest are Machine learning, Information Security, Web Development and Algorithms. He is currently working as Software Developer Engineer at Commonfloor (MaxHeap Technologies).

Ankur Bhardwaj
He has pursued his B.Tech. in Information Technology from National Institute of Technology Karnataka, Surathkal. His areas of interest are Machine Learning, Statistics and Algorithms. He is currently working as Associate Professional at Computer Sciences Corporation.

INTENTIONAL BLANK

A SECURE NAIVE BAYES CLASSIFIER FOR HORIZONTALLY PARTITIONED DATA

Sumana M[1] and Hareesha K S[2]

[1]Department of Information Science and Engineering,
M S Ramaiah Institute of Technology, Bangalore, Karnataka,India
sumana.m@msrit.edu
[2]Department of Computer Applications,
Manipal Institute of Technology, Manipal, Karnataka, India.
hareesh.ks@manipal.edu

ABSTRACT

In order to extract interesting patterns, data available at multiple sites has to be trained. Distributed Data mining enables sites to mine patterns based on the knowledge available at different sites. In the process of sites collaborating to develop a model, it is extremely important to protect the privacy of data or intermediate results. The features of the data maintained at each site are often similar in nature. In this paper, we design an improved privacy-preserving distributed naive Bayesian classifier to train the horizontal data. This trained model is propagated to sites involved in computation. We further analyze the security and complexity of the algorithm.

KEYWORDS

Privacy Preservation, Naive Bayesian, Secure Sum, Classification.

1. INTRODUCTION

Distributed Computing environment allows sites to learn not only its own training dataset but also other sites training datasets. The outcome is considerably better than training at individual sites. Privacy concerns are large when sites collaborate in a distributed system[5,6]. One solution to perform this form of data mining is to have a trusted learner who builds a learning model by collecting all the data from the data holders [5,6,9,10]. However, in many real world cases, it is impossible to locate a trusted learner. Hence this approach is not considered feasible.

Researchers from various sectors such as medical, bank, security systems, finance are keen to obtain the result of cooperative learning without seeing the data available at other parties. For example, three banks in the same city want to know more information about the credit risk evaluation of the customers with the customer information they hold. These banks need to only communicate essential information during the training phase. After the training, the final model is broadcasted to the banks. The customer data held by the individual banks contain lot of private information such as age, marital status, annual wages and amount invested which are protected by law and cannot be revealed without the customer's consent. In another situation, consider a medical research where doctors the different hospitals want to identify whether the right treatment is given for a medical diagnosis without revealing the individual patient's details.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 21–34, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4903

Our solution avoids revealing data beyond its attributes, while still developing a model corresponding to that learned on an integrated data set. Hence we assure that the data maintained at each of the sites are secure. In this paper, we propose privacy preserving Naive Bayesian classifier on horizontally partitioned data maintained at different sites. We handle both numeric and categorical attributes. Our method is based on performing addition using homomorphic encryption technique and uses a secure division protocol on these encrypted values. We have tested our protocol on real datasets.

Our main contributions can be summarized as follows:

- Enhanced privacy while computing the Naive Bayesian Classifier.
- Use of homomorphic property of Paillier cryptosystems to perform Secure sum.
- Use a Secure Division for Numeric and Categorical attributes of the dataset.

Researchers developed protocols to facilitate data mining techniques to be applied while preserving the privacy of the individuals. One approach[1] adds noise to the data before data mining. Agrawal and Srikant[5] proposed data perturbation techniques for privacy preserving classification model construction on centralized data. [1] Discusses building association rules from sanitized datasets. Such methods also called as data distortion methods assume that the values must be kept private from the data mining party. Also obtaining the exact results is a tedious process.

Another form of privacy preservation data mining uses cryptographic techniques to protect privacy. This approach includes secure-multiparty computations to realize perfect privacy. Methods for privacy preserving association rule mining in distributed environments were proposed by Kantarcioglu and Clifton[12]. [7][11][13] Constructs a classifier model using secure multiparty protocols. Classification using neural networks and preserving privacy is discussed in [14][15][16][20]. Another essential data mining tasks developed for privacy preservation has been discussed in [8].

,Kantarcioglu and Vaidya [11] proposed a privacy-preserving naïve Bayes classifier for horizontally partitioned data. For the computation of probability p summations are computed by site1 adding a random number to its value and forwarding it to its neighbour. Other sites add their value to this value and forward it in a circular manner. The first site will interpret the result by subtracting the value received with the random value. Further to obtain the probability = $\sum_{i=1}^{k}$ Pi/$\sum_{i=1}^{k}$ Ci where k is the number of sites. P_i and C_i is the sum of values present at site i is computed by maintaining P_i in site 1 and C_i in last site and using the ℓn() protocol [9]. Though this protocol assumes no collusion among the sites, it is still vulnerable to the eavesdropping attack where any attacker who intercepts all transmissions among all sites is able to derive each site i's secret values. Also this protocol is not suitable if the number of sites n< 3.

In section II we briefly provide the background and related work required to develop our protocol. Section III discusses our algorithm. Security analysis of our protocol is elaborated in section IV.

2. PRELIMINARIES

2.1. Naïve Bayesian Classification

Naïve Bayesian Classifier [11] uses the Bayes Theorem to train the instances in a dataset and classify new instances to the most probable target value. Each instance is identified by its attribute set and a class variable. Given a new instance X with an attribute set, the posterior

probability P(Class1/X),P(Class2/X) etc has to be computed for each of the class variable values based on the information available in the training data. If P (Class1/X)>=P (Class2/X)>=……>=P (ClassN/X) for N class values, then the new instance is classified to Class1or Class2…or ClassN accordingly. This classifier estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label y. The conditional independence can be obtained as follows: $P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y)$, where each attribute set $X = \{X_1,X_2,…..,X_d\}$ consists of d attributes.

Each of the d attributes can be categorical or numeric in nature. Algorithm 1 indicates the computation of the probability for a categorical attribute and Algorithm 2 indicates the computation of mean, variance and standard deviation required for calculating probability.

Algorithm 1 : Handling a categorical attribute

Input: r -> # of class values, p -> #of attribute values
C_{xy} –> represents #of instances having class x and attribute value y.
N_x – > represents # of instances that belong to class x
Output: P_{xy} –> represents the probability of an instance having class x and attribute value y
For all class values y do
{Compute N_x
For every attribute value x
{Compute C_{xy}
Calculate $P_{xy} = C_{xy}/ N_x$}}

Algorithm 2 : Handling a numeric attribute

Input: r -> # of class values, x_{jy} -> value of instance j having class value y.
S_y -> represents the sum of instances having class value y
N_y -> represents # of instances having class value y
For all class values y do
{Compute $S_y = \sum_j x_{jy}$
Compute n_y
Compute Meany = S_y/ n_y
Compute $V_{jy} = (x_{jy} – Mean_y)^2$ for every instance j that belongs to the class y
Compute $Var_j = \sum_j V_{jy}$
Compute $Stan\_dev2_y = Var_j / (N_y-1)$
}

Once the Variance and Standard Deviation is computed the probability for the numeric value provided in the test record for each of the class can be computed as follows:

P (given that (attribute_value = test_record_numeric_value)| Classy)

$$= 1\frac{1}{\sqrt{2\pi}*Stan\ dev}\ \exp-\frac{(test\_record\_numeric\_value-Mean)}{2\ X\ Stan\_dev\ (of\ class\ y)}\ y$$

On obtaining the Probabilities for each of the attributes with respect to each of the classes the class-conditional probabilities can be computed as follows:
For each of the class value I

Probability (test record having z attribute values | classI)= P(Attr1_value|classI)
*P(Attr2_value|classI) *…….* P(Attrz_value|classI)

The test record belongs to the class has the maximum class-conditional probability.

2.2. Paillier Encryption

In our algorithms, a homomorphic cryptographic scheme of Paillier is utilized. This asymmetric public key cryptography [2,18,19] approach of encryption is largely used in privacy preserving data mining methods. The scheme is an additive homomorphic cryptosystem that are used in algorithms where secure computations need to be performed. Paillier is a public key encryption scheme which can be defined on any cyclic group. The original cryptosystem provides semantic security against chosen-plaintext attacks. Let G be a cyclic group of prime order q with generator g.

Key generation

Obtain two large prime numbers p and q randomly selected big integers and independent of each other such that gcd(pq,(p-1)(q-1)) = 1. Compute

n=pq and $\lambda = lcm(p-1, q-1)$

Select random integer g where $g \in Z^*_{n^2}$. Check whether n divides the order of g as follows
Obtain ℓ=((p-1)*(q-1))/gcd(p-1,q-1)

If(gcd(((g^ℓ mod n^2)-1)/n),n)!=1) then select g once again.

Encryption

Encrypts the plaintext m to obtain the Cipher text $c = g^m * r^n \bmod n^2$ .where m plaintext is a BigInteger and ciphertext is also a BigInteger

Decryption

Decrypts ciphertext c to obtain plaintext $m = L(g^{\ell} \bmod n^2) * u \bmod n$, where $u = (L(g^{\ell} \bmod n^2))^{-1} \bmod n$.

Paillier schemes have probabilistic [19] property, which means beside the plain texts, encryption operation needs a random number as input. Under this property there can be many encryptions for the same message. Therefore no individual party can decrypt any message by itself.

2.3. Homomorphic Encryption

Homomorphic encryption[17] is a form of encryption which allows specific computations to be carried out on ciphertext and obtain an encrypted result which decrypted matches the result of operations performed on the plaintext. For instance, one person could add two encrypted numbers and then another person could decrypt the result, without either of them being able to find the value of the individual numbers.

Encryption techniques such as ElGamal[4,2] and Paillier [18] have the homomorphic property i.e for messages m1and m2

$D(E(m1,r1).\ g^{m2}) = m1+m2 \bmod n$ without decrypting any of the two encrypted messages.
Also $D(E(m1,r1)*E(m2,r2) \bmod n^2) = m1 + m2 \bmod n$.
D indicates decryption and E indicates encryption.

2.4. Secure Multiparty Protocols

To solve our problem of secure computation [11] we have used secure protocols for computing the sum and divide. Some of the secure computations have been discussed in [3]. The parties could apply the algorithm to add two values maintained by them without revealing their values to other parties. This protocol has been implemented by utilizing cryptographic schemes with the additive homomorphic property. Secure Division is performed by a single party with the numerator and the denominator in their encrypted form. A detailed description regarding the usage of these protocols is discussed in the next section.

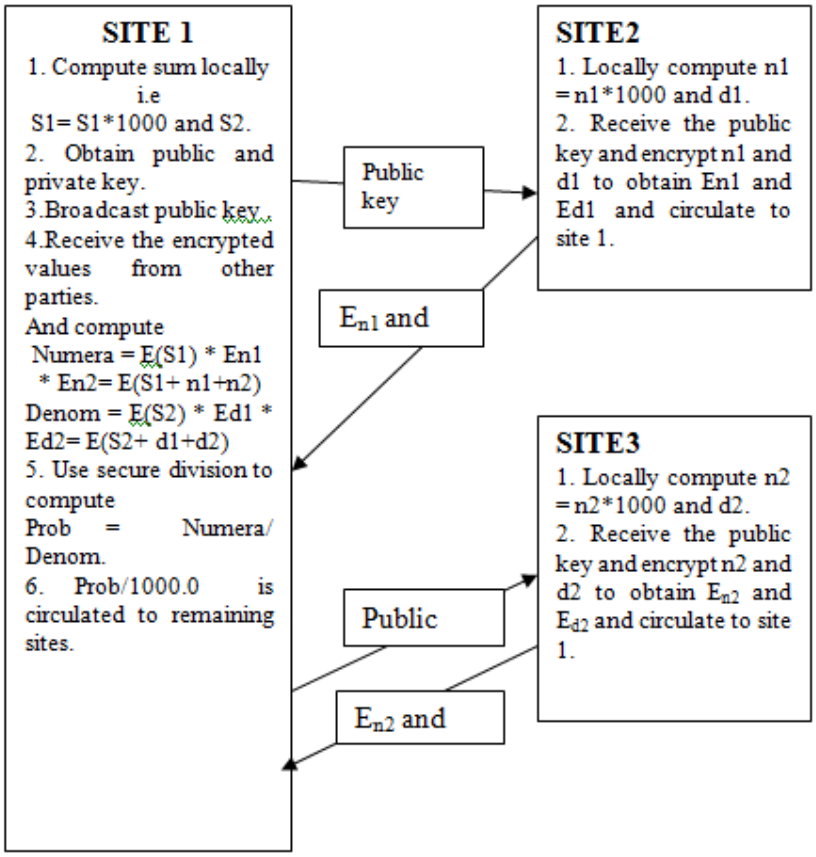
3. MODEL CONSTRUCTION

In this paper we focus on secure training of a horizontally partitioned dataset to build a Naïve Bayesian Classifier model. This constructed allows each of the party to classify a new instance. Multiple banks hold information about the age, Class of Worker , education ,wage per hour, marital stat, major industry code, major occupation code, race, sex ,full or part time employment stat, capital gains, capital losses, dividends from stocks, tax filer stat, region of previous residence, state of previous residence ,detailed household and family stat, num persons worked for employer, family members under 18, country of birth self, citizenship, own business or self employed, veterans benefits, weeks worked in year . This information is collected from people residing in the locality that the banks exist. The characteristics of the individual are either numeric or nominal in nature. Each of the banks has thousands of records holding the information. In order to conclude on a loan decision salary of a person is an important data. Two or more banks want to predict the salary of an individual based on the age, Class of Worker, tax filer status, marital status, qualification, residing region and number of persons in the family. But these banks want to disclose the result of their computation without revealing any other information to a third party or to each other. The above task can be performed by training the horizontally partitioned data in a secure manner.

The protocols presented below are efficient and do not compromise on security. In the process, even the numerator and the denominator of the fractions are not known to any of the parties. Secure Division is performed in a single site. In the following sections we discuss the approaches where only the final classifier is broadcasted to all the parties.

Since all the attributes needed for classifying a new instance are known to all the parties we need not hide any of the attributes or their values. Hence once classifier is given to all the parties, parties need not collaborate to classify a new instance. Also we need not conceal the model. Figure 1 provides a scheme of building the model for 3 parties.

Figure 1: Model Building Process using 3 parties



In this section we discuss the methods for constructing models for both categorical and numerical attributes. Since the procedures for learning is dissimilar for both the types of attributes, we define different methods for each.

3.1. Categorical Attributes

For categorical attributes, the conditional probability has to be computed. Conditional probability gives the probability that an instance belongs to a class 'c' for an attribute A having an attribute value 'a' indicated as$P(C= 'c'/A = 'a')= n_{ac}/n_a$

where n_{ac} – number of instances in the training set(in all the collaborating parties) that have the class value 'c' and value of the attribute value as 'a' and n_a - number of instances(in all the collaborating parties) where attribute A = 'a'.

As the datasets are horizontally partitioned, parties are aware of some or all of the values of their categorical attributes. To compute the sum n_{ac} and n_a for all the parties, each party locally counts the number of instances and then parties collaborate to use the paillier homomorphic secure sum protocol (algorithm 5) to compute the global count. During collaboration, local counts are not revealed to any of the intermediate parties. The party that has initiated the training phase has the encrypted results (both numerator and denominator).These encrypted values are then securely

divided to obtain the probability. As observed in algorithm 3 we multiply the numerator by 1000 and further divide the result with 1000 to round the result obtained by 3 decimal points.

Algorithm 3: Handling a categorical attribute

Input: k parties, r class values, n attribute values
C^i_{ac} – number of instances with party P_i having class c and attribute value a.
n^i_c – number of instances with party P_i having class c.
p_{ac} - Probability of an instance having class c and attribute value a.
for all class value c **do**
for i= 1 to k **do** **// for each party**
for every attribute value a, party P_i locally computes C^i_{ac} .Then Perform $C^i_{ac}= C^i_{ac} * 1000$.
Party P_i locally computes n^i_c // local computation by each party
end for
end for

All parties collaborate using secure sum protocol to obtain $EC_{ac} = E(\sum_{i=1}^{k} C_{i_{ac}})$.

For every class value c, all parties collaborate using secure sum protocol, $En_c = E(\sum_{i=1}^{k} n_{i_c})$.
Party 1 which initiated the model construction computes p_{ac} using the EC_{ac} and En_c using secure division protocol(algorithm 6) . Final $p_{ac} = p_{ac} / 1000$.

3.2. Numeric Attributes

For numeric attributes the mean value has to be securely computed. Mean value of a class 'c' = S_c/n_c , where S_c is the sum of all the instances in the multiple parties belonging to class 'c' and n_c is the number of instances belonging to class 'c'.

All parties at first locally compute the mean of its numeric attribute value. They also obtain the sum of all the instances that belong to the class 'c'. Further algorithm 5 is used to find the encrypted result of the global sum of S_c and n_c. Algorithm 4 is used to give the mean of instances belonging to class 'c'.

Using this mean the parties then collaboratively calculate variance.

Algorithm 4 : Handling a Numeric Attribute

Input : k parties, r class values
x_{icj} - the values of instances j from party i having class value c
s^i_c - the sum of instances from party i having class value c
n^i_c - the number of instances with party P_i having class value c
for all the class values c do
for i= 1 to k do
Party P_i locally computes $s^i_c = \sum_j x_{icj}$. Performs $s^i_c = s^i_c * 1000$.
Party P_i locally computes n^i_c
end for

All parties in collaboration perform secure sum protocol to compute $E(s_c) = E(\sum_{i=1}^{k} s_{i_c})$.
All parties in collaboration perform secure sum protocol compute $E(n_c) = E(\sum_{i=1}^{k} n_{i_c})$

Party 1 computes the mean $\mu_c = E(s_y)/E(n_y)$ using secure division protocol.
mean $\mu_c = \mu_c/1000$;
end for
μ_c is circulated to all the other sites.
//to compute total variance
for i=1 to k do

for every instance j, $v_{icj} = x_{icj} - \mu_c$ and $v_{ic} = \sum_j v2_{icj}$
end for

All the parties then collaborate using secure sum protocol to compute variance

$E(v_c) = E(\sum_{i=1}^{k} v_{ic})$

$D(E(v_c))$ is performed by party 1 to obtain v_c.

Finally party 1 computes stan_dev $\sigma^2_c = \frac{1}{nc-1} v_c$

3.3. Secure Sum Protocol

This algorithm is used to securely compute the sum of the values maintained at individual sites.

Algorithm 5: Secure Sum Protocol

Party P_1 uses randomgenerator to obtain a random number r_1, uses Paillier encryption technique to obtain public key P_k. It uses its public key to encrypt its value S_1 as follows $E(S_1,r_1)$.
for i= 2 to k

Use RandomGenerator to obtain the random number r_i .

Uses the public key to obtain $E(S_i,r_i)$,

and forwards it to party P_1.
end for

Party P_1 then computes Encrypt_prod = $\pi_{k_{i=1}} E(S_i,r_i)$.

Note: $\pi^{k}_{i=1} E(S_i,r_i) = E(S_1+S_2+S_3+\ldots\ldots+S_k)$.

3.4. Secure Division

Since the numerator (n) and the denominator (d) are in the encrypted form we use this method. The encrypted values are of BigInteger type that exceeds the size of 512 bits. The Logic used is the working [23] is as follows:

I. Compute an encrypted approximation [a~] of $a = [2^k/[d]]$

II. Compute [n/d] as $([a\sim]*[n])/ 2^k$.

To compute the k shift approximations of 1/d we use the concept of Taylor's series to define the desired approximation of $2^k/d$ as

$a\sim = 2^{k-\ell d(w+1)} * \sum_{i=0}^{w}(2^{\ell d}-d)i * 2^{\ell d(w-i)}$.

Further we compute using Z_M arithmetic, with M = p * q, which is the Paillier key whose secret key is held jointly by the parties. Hence a~ is modified as follows

$a\sim = 2^{k-\ell d} * \sum_{i=0}^{w}((2^{\ell d}-d)*2^{\ell d})^i$.

Algorithm 6 discusses the secure division protocol. This protocol is being executed at a site with no communication with other parties.

Algorithm 6: Secure Division on encrypted values

Input: Encrypted numerator [n] and encrypted denominator [d](ℓ-bit value)

1. Compute $2^{\ell d}$ from [d]
count =1
obtain binary representation of [d]. Initialize p_0=1
while(count<=$\log_2 \ell$)
begin
c1 =0
if($2^{\ell/2}.p_0$ <= [d])
c1 = 1
$p_0 = p_0*(c1*(2^{\ell/2}-1)+1)$
end
compute $2^{\ell d} = 2*p_0$.

2. Obtain $2^{-\ell d}$ = Inverse($2^{\ell d}$)

3. Obtain Poly (p) for p = ($2^{\ell d}$-d) * ($2^{-\ell d}$) as follows
Use square and multiply method to evaluate $\sum_{i=0}^{w} p^i$ where w = 2^R for some integer R.

4. Compute a~ = 2^k* $2^{-\ell d}$ * Poly (p).

5. Further we find q^ = [n]. a~

6. Truncate q^ by k to acquire q~ is approximately equal to ($q/2^k$) as follows
Obtain [z] -> q^+ r , where r is a random number Є $Z_2^{k+\ell}$.
Decrypt [z] and generate q~ = ($z/2^k$)-($r/2^k$)

7. Eliminate errors generated as follows
r = [n]-[d]* q~
if([d]+[d]<=r+[d])
pos_err = 0.1 else pos_err = 0.0
if([d] >r+[d])
neg_err = 0.1 else neg_err=0.0

8. Finally compute q <- q~ + pos_err + neg_err.

3.5. Classifying an instance

As we have implemented our protocols for horizontally partitioned dataset all the attributes are known to all the parties. The party that wants to evaluate an instance simply uses the probability values obtained for categorical attributes , mean and variance computed for numeric attributes and locally classifies it. It need not interact with the other parties. Hence there is no compromise in privacy.

4. SECURITY ANALYSIS

In this section we elaborate on why our algorithms are secure in the semihonest model. In the semihonest model, the parties during computation are curious and try to analyze the intermediate values and results. Hence in a secure model we must show that the parties learn nothing except their outputs from the information they obtain during the process of execution of the protocol. The encryption scheme, Paillier, used in the protocol is semantically secure as the result each ciphertext can be simulated by a random ciphertext.

Algorithm 1 securely computes the probability pac without revealing anything (i.e. either the global count Cac or global number of instances nc). The only communication occurs while computing the global sum using homomorphic encryption. When there is no collision between the parties each party's view of the protocol is simulated based on its input and its output. Algorithm 2 securely computes the mean μc and variance vc without revealing anything except μc and vc . The communication in this algorithm occurs while computing the global sum while mean and variance but the global sum is not revealed to any of the parties.

In algorithm 1and 2 parties 2 to k communicate with each other with their encrypted values and multiply and forward it to their neighboring party to obtain the encrypted global sum. Party 1 performs an additional step of computing the division of Paillier encrypted values. After the secure division protocol party 1 sees only the result of division which is broadcasted to the other parties.

Even though the public key is known to all the parties and each of the parties encrypt their data to assist in computation because of the probabilistic property of Paillier parties cannot decrypt the other parties' data. Hence we propose that our approach is secure. As mentioned in [23] the secure division protocol does not reveal any information about the inputs(other than the desired encryption of the result).

4.1. Effect of Collusion on Privacy

In our solution, in the process of secure sum additions involving k parties, if C_{ac} and n_c can be evaluated even if k-1 parties collude with each other. However if all of the k parties collude, privacy protection is irrelevant.

For the secure division protocol, since only 1 party performs the computation colluding of the other parties will not affect the protocol. Also if party 1 colludes with the other parties, it only has the encrypted values hence it cannot reveal anything to the other parties.

4.2. Communication and Computation Cost

The secure division protocol requires only O $((\log^2 \ell)(\alpha + \text{loglog}\ \ell))$ arithmetic operations in O $(\log^2 \ell)$ rounds where α is the correctness parameter and ℓ is the size of the numerator and denominator. The computation of $2^{\ell d}$ for encrypted d requires $\log_2 \ell$ iterations, each involving one

comparison and one multiplication. Hence the complexity is O ($\log^2 \ell$). The round complexity of poly (p) is O (log w) where $w = 2^{\lambda}$ approximately equal to ℓ. Further the round complexity of truncating is O (log ℓ).

For calculating conditional probability privately we require k secure additions and one division for k parties. Compared to non-secure version of the conditional probability calculation the secure version is much slower. Computation using homomorphic secure sum protocol involves only k encryptions and k summations; hence the computation cost is dominated by the secure divide protocol. Given a dataset having n_1 categorical attributes with an average of n_a values, the number of global computations performed are $2*(n_{1*} n_a)*$ k secure additions and ($n_{1*} n_a$) secure divisions by party 1. For n_2 numeric attributes, global computations are $3*(n_2)*$k secure additions and n_2 secure divisions by party1. The local computations of sum performed by each of the party's depend majorly on the number of tuples they have. We have implemented our approach with n sites Intel(R) core ™ 2 CPU, 6400 @ 2.13GHz, 2GB ram with a Java program to enable the n sites to interact with each other during secure sum computation.

Given in Table 1 is the computation time for calculating conditional probabilities worked on the census dataset (Salary as class attribute, with occupation, education, marital status, dependency as categorical attributes and age, capital gains as numeric attributes) and breast cancer (Diagnosis as class attribute, all 8 attributes are numeric in nature) from the UCI repository. The table summarizes the approximate computation time for conditional probabilities for different database sizes. The time required to classify a new instance is the same as that in a non-privacy version of the classifier.

For test samples Table 2 indicates the accuracy of our approach. Accuracy is computed as the total number of correctly classified tuples divided by the total number of tuples in test sample.

Table 1: Estimated Computation Time for conditional probabilities

Security Parameter (in bits)	Total tuples in all sites	Degree of the Polynomial	Estimated Time (seconds) Census Dataset	Estimated Time (seconds) Breast Cancer Dataset
512	10^5	10	0.68	0.72
512	10^5	20	0.71	0.75
512	10^6	10	0.84	0.90
512	10^6	20	0.88	0.94
512	10^7	10	0.91	0.96
512	10^7	20	0.95	1.06
1024	10^5	10	2.75	2.83
1024	10^5	20	2.86	2.92
1024	10^6	10	3.51	3.69
1024	10^6	20	3.72	3.81
1024	10^7	10	4.56	4.62
1024	10^7	20	4.64	4.78

Table 2: Accuracy of the classifier

Size of test samples	Accuracy(%)
10^3	83
10^4	85
10^5	87

4.3 Implementation

The algorithms are implemented in Java in Eclipse IDE. The testing data sets are from the Irvine dataset repository. We choose the census data set where we use 14 categorical and 7 numeric attributes for building a model on the salary class attribute. We have performed experiments based on the varied size of the datasets maintained at other parties.

Experimental Results

We have performed our experiments on the non-privacy naïve Bayesian classification version the privacy versions that we have implemented. We calculate the Classifier Accuracy = (Number of test samples misclassified)/(Total number of samples).For the census dataset with the salary attribute as class label attribute our results is mentioned in Table 3. Our approaches are quite effective in learning real world datasets. Also cryptographic algorithms are essential whenever there are privacy issues.

Table 3 : Accuracy of classifiers

Algorithm	**Accuracy(approx)**
Non-privacy Naïve Bayesian	85%
kantarcioglu-Vaidya's Privacy Preserving Naïve Bayesian Approach[11]	78%
Our Privacy Preserving Naïve Bayesian on Horizontally Partitioned Data	83%

As observed in Table 3 our approach provides accuracy nearly as that of the Non-privacy Naïve Bayesian on our distributed dataset maintained at sites greater than 3. Though the computation time required for our classifier is more than the non-privacy version as well as kantarcioglu-Vaidya's Privacy Preserving Naïve Bayesian Approach[11], our classifier is more secure(since encryption is used) and provides better accuracy on our data set.

5. CONCLUSION

This paper concentrates on building a secure Naïve Bayesian classifier with multiple parties without revealing any information during summation and division. The probability, mean and variance obtained securely are circulated to all the parties for classifying a new instance.

Our approach even though expensive than the non-privacy version of the protocol thrives to achieve a model that is secure and efficient. The algorithm guaranteed privacy in a standard

cryptographic model, the semi honest. In future we intend to explore privacy preservation approaches for other classifiers.

ACKNOWLEDGEMENTS

We would like to thank M S Ramaiah Institute of Technology and Manipal Institute of Technology for their constant support in our research work.

REFERENCES

[1] L. Wan,W.K.Ng,S.Han, and V.C.S.Lee,"Privacy Preservation for gradient descent methods,"in Proc.ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2007,pp.775-783.

[2] O.Goldreich, "Foundations of Cryptography".Cambridge, U.K.:Cambridge Univ. Press, 2001-2004, vol. 1 and 2.

[3] A.Yao, "How to generate and exchange secrets," in Proc. 27th IEEE Symposium Foundations, Computer Science., 1986, pp. 162-167

[4] T.ElGamal,"A public-key cryptosystem and a signature scheme based on discrete logarithms," IEEE Transactions of Information Theory, vol.IT-31, no.4, pp.469-472, Jul.1985.

[5] D.Agrawal and R.Srikant,"Privacy-preserving data mining," in Proc. ACM SIGMOD, 2000, pp.439-450.

[6] Y.Lindell and B.Pinkas, "Privacy preserving data mining," in Lecture Notes in Computer Science, Berlin, Germany: Springer-Verlag, 2000, pp.36-44.

[7] N.Zhang, S.Wang, and W.Zhao, " A new scheme on privacy-preserving data classification, " in Proc. ACM SIGKDD International Conference of Knowledge discovery and data mining,2005,pp.374-383.

[8] G. Jagannathan and R.N. Wright,"Privacy-preserving distributed k-means clustering over arbitrarily partitioned data," In Proceedings of ACM SIGKDD International Conference of Knowledge discovery and data mining 2005,pp 593-599.

[9] Y.Lindell and B.Pinkas, "Privacy preserving data mining," Journal of Cryptography, volume 15, no 3,2002, pp177-206.

[10] R Agrawal , R Srikant: "Privacy-preserving data mining" In Proceedings of the 2000 ACM SIGMOD Conference on Management of Data, pp.439-450. ACM, Dallas, TX(2000). http://doi.acm.org/10.1145/342009.335438.

[11] J Vaidya.,M Kantarcioglu,C Clifton.,: Privacy Preserving Naïve Bayes Classification. In: The VLDB Journal (2008) 17: 879-898, DOI 10.1007/s00778-006-0041-y.

[12] M Kantarcioglu, C. Clifton, Privacy-preserving distributed mining of association rules on horizontally partitioned data, IEEE TKDE 16(9) (2004).

[13] S. Samet, A. Miri, Privacy Preserving ID3 using Gini Index over horizontally partitioned data ,In: Proceeding of the 6th ACS/IEEE International Conference on Computer Systems and Applications(AICCSA), Doha, Qatar, 2008,pp.645-651.

[14] R. Wright and Z. Yang , " Privacy-preserving Bayesian network structure computation on distributed heterogeneous data," in Proc. 10th ACM SIGKDD Int. Conf.knowledge.Disc. Data Mining, 2004,pp.713-718.

[15] A. Bansal, T Chen, S Zhong., Privacy Preserving Back-propagation neural network learning over arbitrarily partitioned data, Neural Computing and Appications(2011) 20: 143-150.

[16] T. Chen and S Zhong., Privacy-Preserving Backpropagation Neural Network Learning, IEEE Transactions on Neural Networks, Vol., 20, No10, 2009.

[17] Mitchell, T.: Machine Learning, 1st edn. McGraw-Hill Science/Engineering/Math, New York(1997).

[18] P.Paillier, Public-key cryptosystems based on composite degree residuosity classes, In: Proceedings of the International Conference on the Theory and Application of Cryptographic Techniques, Prague, Czech Republic,1999,pp. 223-238.

[19] Blum, M., Goldwasser, S.: An efficient probabilistic public-key encryption that hides all partial information. In: R. Blakely(ed.) Advances in Cryptography-Crypto 84 Proceedings, Springer, Heidelberg(1984).

[20] S. Samet, A.Miri, Privacy Preserving back-propagation and extreme learning machine algorithms, Data and Knowledge Engineering, 79-80(2012) 40-61

[21] Blake CL, MErz CJ(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.
[22] O Goldreich " The Foundations of Cryptography", vol.2, chap. General Cryptographic Protocols. Cambridge University Press, Cambridge University Press, Cambridge (2009).
[23] M Dahl, Chao Ning, T Toft, "On Secure Two-party Integer Division", Financial Cryptography and Data Security ,Lecture Notes in Computer Science, Volume 7397, 2012, pp 164-178.

AUTHORS

Sumana M was born in Madikeri, Karnataka, India on 15th December 1978. She has received her B.E. degree in Computer Science and engineering from Manipal Institute of Technology, Karnataka, India, in 2000, and her M.Tech. degree from VTU University in 2007, Karnataka, India and is currently pursuing Ph.D. degree in privacy preserving data mining in computer science and Engineering from the Manipal University, Karnataka, India.

She is presently working as an Assistant Professor in the department of Information Science and Engineering in M S Ramaiah Institute of Technology since 2007. Previously she had worked as a lecturer in the Manipal Institute of Technology. Her research interests include data mining, cryptography and secure multiparty computations. She is a Life Member of the Indian Society for Technical Education (ISTE), the System Society of India.

Hareesha K.S. born in Chikmagalore, Karnataka State, India on 11th April, 1970. He has received the BCA, MCA in computer applications and Ph.D. in computer science & engineering from Kuvempu Universtiy, Shankaraghatta, Karnataka, India in the year 2000, 2003 and 2008 respectively. His major research interest is digital image processing, data mining, bioinformatics and artificial intelligence.

Since 2008, he is working in Manipal University, Karnataka India as Associate Professor, before to this he was worked in Bapuji Institute of Technology, Davangere, Karnataka, India. Also, took charge as Head of Department, Department of Computer Applications, Manipal Institute of Technology, Manipal University from Jan. 2013. His research interests encompass privacy preservations in data mining, spatial data mining and its relevance in society as a whole, bio-informatics, biologically inspired algorithms, soft computing and computer vision. He has published quite a good number of research papers in these areas. He has got fellowship award from Boston University to present his research contributions. Dr. Hareesha K.S. has been a life member of ISTE, New Delhi, India and IACSIT, Singapore. He has been a member of numerous program committee of IEEE, IACSIT conferences in the area of Data Mining, Bioinformatics, Digital Image Processing and Artificial Intelligence.

ONTOLOGY BASED DATA MINING METHODOLOGY FOR DISCRIMINATION PREVENTION

Nandana Nagabhushana[1] and Dr.Natarajan S[2]

[1]M.Tech in Software Engineering,
Department of Information Science, PESIT, Bangalore, India
nandana.sgn@gmail.com
[2]Professor and Key Resource person,
Department of Information Science, PESIT, Bangalore, India
natarajan@pes.edu

ABSTRACT

Data Mining is being increasingly used in the field of automation of decision making processes, which involve extraction and discovery of information hidden in large volumes of collected data. Nonetheless, there are negative perceptions like privacy invasion and potential discrimination which contribute as hindrances to the use of data mining methodologies in software systems employing automated decision making. Loan granting, Employment, Insurance Premium calculation, Admissions in Educational Institutions etc., can make use of data mining to effectively prevent human biases pertaining to certain attributes like gender, nationality, race etc. in critical decision making. The proposed methodology prevents discriminatory rules ensuing due to the presence of certain information regarding sensitive discriminatory attributes in the data itself. Two aspects of novelty in the proposal are, first, the rule mining technique based on ontologies and the second, concerning generalization and transformation of the mined rules that are quantized as discriminatory, into non-discriminatory ones.

KEYWORDS

Ontology, Discrimination Prevention, Rule Protection, Rule Generalization, Postmining

1. INTRODUCTION

The unjust or prejudicial treatment of different categories of people, especially on the grounds of race, age, or gender is coined as Discrimination. It is the recognition and understanding of the difference between one quality and another, which might pave way for inequity and bigotry towards some particular classes of society in provision of certain services, which otherwise should be made obtainable to all the classes of the society. The Anti-Discrimination Acts proposed and institutionalized as a part of Law of the land by various nations, consist several clauses designed to prevent discrimination in numerous fronts like access to public services, loans, insurance, education, employment etc. based on attributes related to Gender, Nationality, Race, Religion, Marital Status, Physical Disability etc. Technology, particularly data mining can contribute to a fair extent in this arena, to discover and prevent discrimination by automating the routines used in many systems for decision making. Collections of data can be used to train association/classification rules to make decisions that are not influenced by the human decision maker who can be probably biased.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 35–48, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4904

Nevertheless, this is not sufficient to abrogate the plausibility of discrimination. A thoughtful contemplation about the data mining process reveals that rules indeed are mined and learnt from a training data-set, which can be inherently biased. This will lead to discovery of rules which are naturally prejudiced, and possibly discriminatory, thereupon necessitating the extermination of potential biases from the training dataset, thus preventing data mining itself being an agent for discrimination.

A novel solution to the problem has been suggested by Sara Hajian et al. [1] for all discovered types of discrimination. Discrimination can be classified as Direct and Indirect/Systematic [2], based on the nature of discrimination implication. Direct Discrimination can be defined as the process of differentiating based on evidently discriminatory attributes related to a disadvantaged group possessing sensitive discriminatory attributes. For example, denial of admission to an educational institution, based on the candidate's ethnicity, can be termed as Direct Discrimination. Indirect Discrimination is differentiation based on certain attributes of the individual that apparently are not discriminatory, but are highly correlated to discriminatory attributes. To exemplify, denial of admission to an educational institution based on the zip-code of the candidate due to the background knowledge that the dwelling of the candidate is mostly occupied by a particular ethnic group.

Drifting towards the technical aspects of the proposal, it is approving to mention that Association Rule Mining forms the backbone of Knowledge Discovery Process. But it is conspicuous that though rule mining aims at discovering implicative tendencies in the collected data, which can be valuable in decision making. It yields to rules whose usefulness is greatly influenced and limited due to their large numbers. Thus, an effort is required to be made to moderate the number of rules learnt from the training dataset.

Based on the literature in [3], a methodology is proposed to conceptualise the background knowledge possessed by the user, in the form of ontologies. Ontologies are constructed and used to formulate and mine rules into the rule schema, which are then subjected to certain transformations. As the last step, an attempt is made to quantize the discrimination present in the final set of rules, and these rules are validated against certain metrics. The rules which pass the threshold test are marked and allowed as non-discriminatory rules, which are collated as the final rule set.

1.1 Related Work

Data Mining has been extensively employed in numerous applications of various domains which inculcate decision making processes. T. Delenius [4] was the harbinger, who in 1970s, first studied and formulated the statistical disclosure control problem. Research has been carried on ever since then, and in 1990s k-anonymity model was proposed by P. Samarati and L. Sweeney [5]. In this approach, a data set is k-anonymous if its records are not distinguishable by an intruder within groups of k members. The novelty of this model was that the anonymity target was established ex-ante and then computational procedures were used to achieve that target.

Decision Models are mostly constructed by machine learning that happens on historical decision records, using data mining methods. Nevertheless, there is no recognizance that automation of decision making completely rules out the chances of production of discriminatory rules, because the extracted knowledge might contain implicit discriminatory bias. An upright approach to prevent this, is to avoid the classifier's prediction to be based on discriminatory attributes by removing them. But, research by F. Kamiran and T. Calders [6] has proved that this is not an effective and efficient method for discrimination prevention. The attributes which highly correlated to the discriminatory attribute can still exist, whose removal might cause information

loss, leading to sub-optimal predictors as depicted in [6, 7].

In this accord, researchers have formalised many strategies among which three are popularised, and practised. The first approach is based on pre-processing, in which the data set is transformed so that the discriminatory rules do not ensue from mining. Kamiran and T. Calders[8, 9] have adopted hierarchy based generalization, to perform controlled distortion and learn the classifier by minimally intrusive modifications to the data sets. This results in an unbiased data set which can then be used to learn rules that are non-discriminatory. This approach proves to be useful in scenarios where the data sets should be published. The in-processing strategy states certain modifications on the data mining algorithms. A novel in-processing method proposed in [10], states that the non-discriminatory criteria are considered as the splitting criteria of a decision tree learner and relabeling is used for pruning. The third strategy, being the post-processing approach, proposes to modify the resulting data-mining model. That is, the rules that are mined as a result of learning the dataset are transformed to remove discrimination. D. Pedreschi, S. Ruggieri, and F. Turini [2, 11] propose a confidence altering approach on the CPAR algorithm. A more recent methodology by Sara Hajian et al. [1], proposes a unified approach to direct and indirect discrimination and also states utility measures to quantify the discrimination. Data transformation methods like rule-generalization and rule-protection are formulated.

In [12], the authors describe and adopt a discrimination discovery method, that not only addresses direct discriminatory attributes, but also those correlated indirect discriminatory attributes. The correlation information is implied as background knowledge, which takes the form of a set of association rules. The challenge of representing the user knowledge has been addressed in a novel way by Claudia Marinica and Fabrice Guillet [3]. It has been proposed that, ontologies can be formalized using specification languages, which can be understood by machines, and parsed in software programs. As a base to this proposal, Liu et al. [3] has proposed a specification language, which can be used to formalize ontologies. In [14], T.R. Gruber defines ontology as a formal, explicit specification of a shared conceptualization. It can be presumed that ontology describes an abstract model of some phenomenon by its important concepts. Also, the formal notion denotes that the formulation and representation of ontology is such that, it is machine interpretable. H. Nigro et al. [15] have classified ontologies into two qualitative categories - Domain and Background Knowledge Ontologies, and, Ontologies for Data Mining Process or Metadata Ontologies.

1.2 Contribution and Plan of this paper

Despite the fact that there have been many propositions of discrimination prevention methodologies, this avenue provides a greater scope for exploration. In this direction, this paper makes an effort to propose a data mining methodology for discrimination prevention using ontologies. This is believed to help in construction of background knowledge by design and offer native technological safeguard against discrimination. This is an attempt to go beyond discrimination discovery and prevention, and cope to the more challenging goal of preventing discrimination in the early stages of KDD process.

This paper is structured as the following: Section 2 introduces notations and definitions used throughout the paper. Section 3 presents the proposed framework and its elements. Section 4 is devoted to the results obtained during experimentation. Finally, Section 5 presents conclusions and shows directions for future research.

2. NOTATIONS AND DEFINITIONS

Let $I = \{i_1, \ldots, i_n\}$ be a set of items, where each item ij has the form attribute=value (e.g.,

Sex=female).

An item set $X \in I$ is a collection of one or more items, e.g. {Sex=female, Credit history=not-taken}.

A ***database*** is a collection of data objects (records) and their attributes; more formally, a (transaction) database $D = \{r_1,...,r_m\}$ is a set of data records or transactions where each $r_i \subset I$. Alternately, the database D can also be defined as a set of transactions $D = \{t_1,...,t_m\}$. Civil rights laws [6, 22] explicitly identify the groups to be protected against discrimination, such as minorities and disadvantaged people, e.g., women.

In the project context, these groups can be represented as items, e.g., Sex=female, which we call Potentially Discriminatory (PD) items; The discrimination is evident with respect to such attributes. A collection of PD items can be represented as an itemset, e.g., {Sex=female, Foreign worker=yes}, which we call PD itemset or protected by-law groups, denoted by DI_s.

An ***itemset*** X is Potentially Non-Discriminatory (PND) if $X \cap DI_s = \emptyset$, e.g. {credit history=no-taken} is a PND itemset where DIs : {Sex=female, Race=black, Foreign worker=yes}.

A ***decision attribute*** is an attribute which takes as values "yes" or "no" to report the outcome of a decision made on an individual. An example for this type of attribute is "credit approved", which can be yes or no. A class item is an item of class attribute, e.g., Credit approved=no.
The ***support*** of an itemset X in a database D is the number of records that contain X. That is, $suppD(X) = | r_i \in D | X \subseteq r_i \} |$, where $| . |$ is the cardinality operator.

An ***Association Rule*** is an implication $X \rightarrow Y$, where X and Y are itemsets and $X \cap Y = \emptyset$. The former is the antecedent and the latter is the consequent of the rule. $X \rightarrow Y$ is a classification rule if Y is a class item and X is an itemset containing no class item e.g. {Sex=female, Credit history=not-taken → Credit approved=no}. The itemset X is called the premise of the rule.

The rule $X \rightarrow Y$ is ***completely supported*** by a record if both X and Y appears in the record. Henceforth, due to generalization of the measures to the context of the considered database, this context suffix in discarded and generalized measures and rules are used.
The ***confidence*** of a classification rule, $conf(X \rightarrow Y)$, is the measure of frequency of the class item Y in records that contain X. Hence, if $supp(X) > 0$ then,

$$conf = \frac{supp(X,Y)}{supp(X)} \quad \text{.....(1)}$$

The value of confidence ranges over [0, 1]

The ***lift*** of a classification rule $lift_D(X \rightarrow Y)$, is the measure of importance of the rule. The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule.

$$lift_D = \frac{conf(X,Y)}{expected\_conf(X,Y)} \quad \text{.....(2)}$$

The ***expected confidence*** of a rule is defined as the product of the support values of the rule antecedent and the rule consequent divided by the support of the rule antecedent.

$$\textbf{expected\_conf}_D(X,Y) = (\textbf{supp}_D(X) * \textbf{supp}_D(Y)) / \textbf{supp}_D(X) \quad \ldots\ldots(3)$$

A ***frequent classification rule*** is a classification rule with support and confidence greater than respective specified lower bounds.

A ***negated itemset***, i.e. ¬X is an itemset with the same attributes as X, but the attributes in ¬X take any value except those taken by attributes in X. For a binary attribute, e.g. {Foreign worker=Yes/No}, if X is {Foreign worker=Yes}, then ¬X is {Foreign worker=No}. For a non-binary categorical attribute, e.g. {Race=Black/White/Indian}, if X is {Race=Black}, then ¬X is {Race=White} or {Race=Indian}. In the current context, only non-ambiguous negations are used.

A ***closed itemset*** [16] is defined as an itemset X which has the property of being the same as its closure, i.e., $X = c_{it}(X)$. The minimal closed itemset containing an itemset Y is obtained by applying the closure operator cit to Y. Let R_1 and R_2 be two association rules. We say that rule R_1 is more general [3] than rule R_2, denoted$R_1 \leq R_2$, if R_2 can be generated by adding additional items to either the antecedent or consequent of R_1. In this case, we say that a rule R_j is redundant [17] if there exists some rule R_i such that $R_i \leq R_j$.

Formally, an ***Ontology*** [18] is a quintuple O = {C, I, R, H, A}. C = {C1, C2,..., Cn} is a set of concepts and R = {R1, R2,…,Rm} is a set of relations defined over concepts. I is a set of instances of concepts and H is a Directed Acyclic Graph (DAG) defined by the subsumption relation (is-a relation, ≤) between concepts. We say that C2 is-a C1, C1 ≤ C2, if the concept C1 subsumes the concept C2. A is a set of axioms bringing additional constraints on the ontology.

3. DISCRIMINATION PREVENTION USING ONTOLOGIES : APPROACH

The dataset used in the case study is titled “Adult Data set” which was extracted by Barry Becker from the 1994 Census database. It comprises of 48842 instances of 14 attributes of type either categorical or integer. Some of the important attributes are age, education, race, sex, and native-country.

The proposed approach can be paraphrased in four phases namely –

1. Ontology construction and rule mining
2. Discrimination measurement
3. Data Transformation

The description of each of these phases follows in the sections 3.1 through 3.3 respectively.

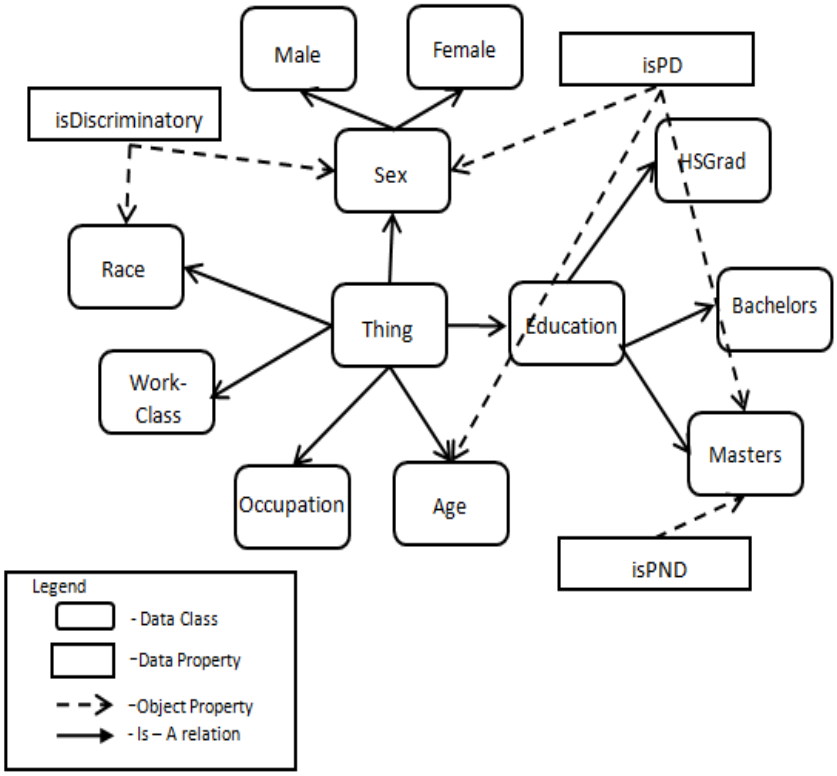


Figure 1: Graphical Representation of the Ontology for Adult Data Set Attributes

3.1 Construction of ontology based on the background knowledge and Association Rule Mining

Background knowledge represents the backbone of association/classification rule mining systems. It is proposed here that ontologies can contribute to a major extent in representing this knowledge. Generally ontologies represent subsumption relations (is-a). The proposal here is to represent background knowledge in the ontology in terms of relationship classes by defining data properties pertaining to discrimination prevention. That is, to represent the knowledge of PD, PND attributes and the subsumption attributes in the ontology. In this accord, four data properties –

- isDiscriminatory,
- isNonDiscriminatory
- isPotentiallyDiscriminatory
- isPotentiallyNonDiscriminatory

are defined in the ontology. The illustration is shown in the Figure 1, which is the representation of ontology construction for the attributes of Adult Data Set.

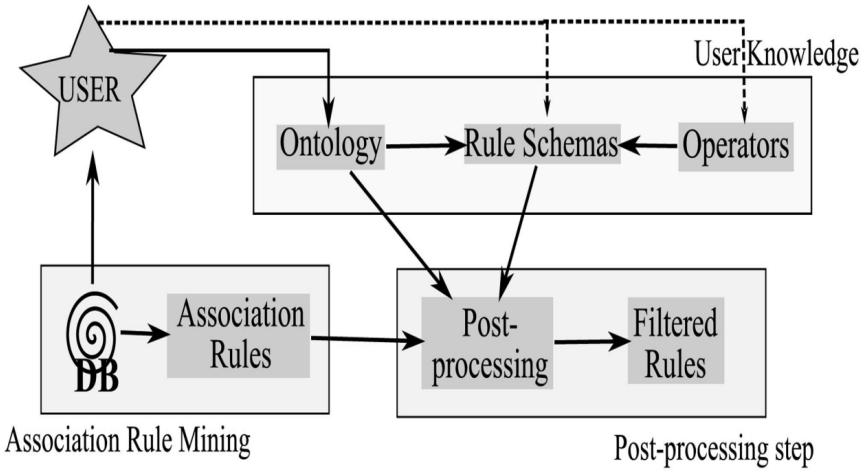


Figure 2: The ARIPSO framework [3]

The ARIPSO (Association Rule Interactive post-Processing using Schemas and Ontologies) framework [3] is used to learn the association rules. Shown in Figure 2 is the ARIPSO framework. One iteration of this process is used, against the suggestion of multiple iterations of user feedback. This is due to the assumption that most of the user knowledge is represented in the data properties of the ontology at one shot. Concepts like, interestingness measures, ontology based rule mining and filtering- results in comparatively less number of rules – rules that are interesting and relevant to the context. The ARIPSO framework chooses to employ FP-Growth algorithm to mine frequent itemsets and hence a set of association rules pertaining to the dataset.

3.2 Discrimination measurement

Although discrimination is discovered in terms of background knowledge during the rule learning phase, a reiteration of this activity is necessary to further classify and fine tune the discovered rules. The utility measures described by Pedreschi et al.[2, 20] over classification rules, for measuring the degree of discrimination of a PD rule (i.e. elift) for direct discriminatory discovery and a PND rule (i.e. elb) for indirect discrimination can be well utilized to quantize the amount of discrimination in each of the generated rule. Filtering of rules should be done based on the threshold values of these measures which further reduces the number of rules. A brief formal description of the terminology and the utility measures follows –

Let, DIs be the set of predetermined discriminatory items in DB (e.g. DI_s = {Foreign worker=Yes, Race=Black, Gender=Female}). Frequent classification rules fall into one of the following two classes:

1) A classification rule X → C is potentially discriminatory (PD) when X = A, B with A ⊂ DI_S, a non-empty discriminatory itemset and B a non-discriminatory itemset. For example {Foreign worker=Yes; City=NYC} → Hire=No.
2) A classification rule X → C is potentially non-discriminatory (PND) when X = {D, B} is a non-discriminatory itemset. For example {Zip=10451, City=NYC} → Hire=No, or {Experience=Low; City=NYC} → Hire=No.

elift is a measure that can be used to assess whether the PD rule is potentially directly discriminatory. Based on a fixed threshold of this measure, a PD rule is judged to be either discriminatory or protective. Formal definition of elift is –

If A,B→ C is a classification rule such that conf(B→ C) > 0, extended lift of the rule is

$$elift(A,B \rightarrow C) = \frac{conf(A,B \rightarrow C)}{conf(B \rightarrow C)} \quad \ldots\ldots(4)$$

where, A ⊂ DI_S and B ∩ DI_S= Ø

Theoretically, elift is the evaluation of discrimination of a rule as a gain of confidence due to the presence of discriminatory items in the antecedent of the rule. If α ∈ R is a fixed threshold stating an acceptable level of discrimination, and if A ⊂ DI_s, and B ∩ DI_s= Ø, then a PD classification rule R1: A, B→ C is α-protective w.r.t. elift if, elift(R1) < α, otherwise it is α-discriminatory.

The PND counterpart of elift is ***elb*** which is used to assess the quantization of discrimination in PND rules. Based on this measure, PND rules can be classified as either redlining or non-redlining (legitimate) rules. To determine the redlining rules, the value of elb is formally arrived at, by the following theorem which provides a lower bound for α-discrimination, using the information available in PND rules which are (γ, δ) and the information (β1 ,β2) available from background rules. The assumption is that the background knowledge takes the form of association rules relating a PND itemset D to a PD itemset A within the context B.

Let r: D, B→ C be a PND classification rule. Let γ = conf(r: D, B→ C) and δ = conf(r:B→ C) > 0 . Let A be a PD itemset, and let β1, β2 be such that,

$$conf\ (r_{b1} : A, B \rightarrow D) \geq \beta 1$$
$$conf\ (r_{b2} : D, B \rightarrow A) \geq \beta 2 > 0$$

Then,

$$f(x) = \frac{\beta 1}{\beta 2}(\beta 2 + x - 1)$$

$$\mathbf{elb(x,y)} = \begin{cases} \frac{f(x)}{y} \ \boldsymbol{iff(x) > 0; 0\ otherwise} \end{cases} \quad \ldots\ldots(5)$$

It holds that, for α ≥ 0, if elb (γ, δ) ≥ α, the PD classification rule R1: A, B→C is α-discriminatory.

A PND classification rule r: D, B→ C is a redlining rule if it could yield an α - discriminatory rule r′: A, B→C in combination with currently available background knowledge rules of the form r_{b1}: A, B→D and r_{b2}: D, B→A, where A is a discriminatory item set. For example, {Zip = 10451; City = NYC} → Hire = No. Otherwise, it is a non-redlining or legitimate rule. For example, {Experience = Low; City = NYC} → Hire = No.

3.3 Data Transformation

Sara Hajian et al. [1] have proposed two data transformation methods – Rule Protection and Rule Generalization which when applied, transforms the data, with minimum information loss. The α – discriminatory rules are transformed to α – protective for direct discrimination, and to an instance of non-redlining PND rule in the case of indirect discrimination.

3.3.1 Rule protection

Rule protection for direct discrimination is termed as Direct Rule Protection (DRP) and is based on the direct discriminatory measure elift. This method simply states that the α – discriminatory rule after transformation, should exhibit an elift less than the value of α. That is if r′: A, B→ C is the transformed counterpart of the rule r, then

$$\textbf{elift } (r') < \alpha \quad \text{.....(6)}$$

From equation 4, we can deduce that

$$\frac{conf(r' : A, B \rightarrow C)}{conf(B \rightarrow C)} < \alpha \quad \text{.....(7)}$$

Thus, by inferring from equation 1, we can achieve the inequality by performing the transformation as stated in Table 1, for the measure elift. This method is a modified version of confidence altering approach stated in [11]. The DRP data transformation attempts to alter the confidence of the base rule B→ C.

On the same lines of DRP, but with the discriminatory measure elb, for indirect discrimination, the transformations are as stated for elb measure in Table 1. The inequality to be established for each redlining rule r: D, B→ C is

$$\textit{elb } (\gamma, \delta) < \alpha \quad \text{.....(8)}$$

The inference for this transformation is from equation 5. These transformations are elaborated and proved in [1].

Table 1: Rule Protection

Measure	Transformation	Condition
elift	¬A, B→ ¬C » A, B→ ¬C	elift(A, B→ C)<α
	¬A, B→ ¬C » ¬A, B→ ¬C	
elb	¬A, B, ¬D → ¬C » A, B, ¬D → ¬C	elb (γ, δ) < α
	¬A, B, ¬D → ¬C» ¬A, B, ¬D →C	

3.3.2 Rule generalization

After performing rule protection, there might still exist some discriminatory content in the rule repository. Unlike the strategies suggested by Pedreschi et al.[12] and by Sara Hajian et al. [1], a simpler method of generalization which makes use of k-anonymity principle proposed and extended by P. Samarati and L. Sweeney[5, 20, 21] is employed. The recourse from the basic k-anonymity theory is, only those α – discriminatory rules that are not subjected to and remain after rule protection, are generalized by anonymization of the PD attribute to the level in the class hierarchy until the rule becomes α – protective or non-redlining. The graph denoted by Figure 3 is an example of the data classification hierarchy for an attribute "Race" in the Adult Data Set. Likewise, if any rule R1: {Race = Australian-White, Age = Young} is generalised to one level higher in the class hierarchy and measured for its discrimination,

Table 2: Adult Data Set Hierarchies

Attribute	No. of Distinct Values	Levels of Hierarchy
Education	16	5
Marital status	7	4
Native country	40	5
Occupation	14	3

Race	5	3
Relationship	6	3
Sex	2	2
Work-class	8	5

and "Hire = No" proves to be α – discriminatory, then it is transformed to R1′ : {Race = White, Age = Young} → Hire = No . If this rule exhibits high values of elift or elb, generalization is reiterated and the rule becomes R1″ : {Race = Any, Age = Young} → Hire = No. Since information loss is inherent with data transformations, the effect of data transformation on data quality should be quantified and measured. Two metrics have been proposed in the literature as information loss measures in the context of rule hiding for privacy-preserving data mining (PPDM) [19] namely Misses Cost and Ghost Cost can be effectively used for this purpose. ***Misses cost*** (MC) quantifies the percentage of rules among those extractable from the original data set that cannot be extracted from the transformed data set. ***Ghost cost*** (GC) is the measure that quantifies the percentage of the rules among those extractable from the transformed data set that were not extractable from the original data set. Generalizarion is performed on all the attributes listed in Table 2. The hierarchy for each of the attributes is obtained from [22]. Additionally, four utility measures [1] have been adopted to measure the discrimination removal. They are –

1. ***Direct Discrimination Prevention Degree*** (DDPD) – Quantifies the percentage of α – discriminatory rules that are transformed into α –protective rules, after the transformations

2. ***Direct Discrimination Protection Preservation*** (DDPP) – Quantifies the percentage of α–protective rules that remain α-protective, after the transformations

3. ***Indirect Discrimination Prevention Degree*** (IDPD) – Quantifies the percentage of redlining rules that transformed to non-redlining, after the transformations

4. ***Indirect Discrimination Protection Preservation*** (IDPP) – Quantifies the percentage of non-redlining rules that remain non-redlining, after the transformations

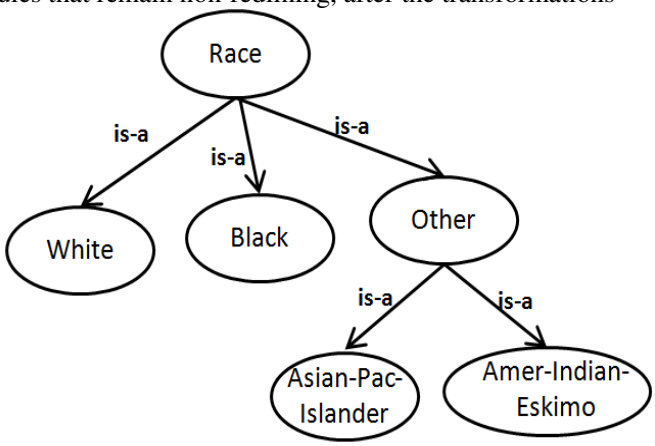


Figure 3: Class Hierarchy for attribute "Race"

4. RESULTS AND ANALYSIS

This section presents the experimental evaluation for the proposed discrimination prevention

system using ontologies. The algorithms were implemented using Java programming language. The ontology and hierarchy graphs have been created using protégé 4.3.0 tool, which is a collaborative development effort between Stanford University and University of Manchester. The tests were performed on a 2 GHz Intel Core i7 machine, equipped with 4 GB of RAM, and running under 64 bit Windows 8 Operating System.

The proposed method for Discrimination Prevention was implemented and evaluated in terms of utility measures. Table 3 shows the results of direct and indirect discrimination prevention for 5% confidence and 10% support at three different levels of α. Since the number of rules generated is considerably low in the case of ARIPSO framework, as depicted by Figure 4, the computational cost proportionally decreases. Table 4 shows the comparison between the direct and indirect discrimination prevention method [1] here after referred as method-1 and the proposed method here after referred as method-2, which happens to be a modified evolution of method-1. These results are based on DI_s = {Foreign worker=Yes, Race=Black, Gender=Female} for rule protection, and all the attributes listed in Table 2 for rule generalization. In these tables "n.a." implies that the respective metrics are not applicable for that method.

Table 3: Utility Measures at Support = 5% and Confidence = 10% on Adult Data Set

α	No. of Rules	No. of α-discriminatory rules	No. of redlining rules	DDPD	DDPP	IDPD	IDPP	MC	GC
α =1	238	38	23	92.2	n.a.	88.3	n.a.	9.4	n.a.
α =1.5	193	29	17	94.5	n.a.	91.1	n.a.	22	n.a.
α =2	167	22	9	95.1	n.a.	92.8	n.a.	27.3	n.a.

Table 4: Utility Measures at Support=5%, Confidence=10% and α =2 on Adult Data Set

Method	No. of Rules	No. of α-discriminatory rules	No. of redlining rules	DDPD	DDPP	IDPD	IDPP	MC	GC
Method-1	204	31	15	93.47	100	~93	100	15.24	4.7
Method-2	167	22	9	95.1	n.a.	92.8	n.a.	27.3	n.a.

The comparison of both the methods against all the considered utility measures is summarized in table 4.

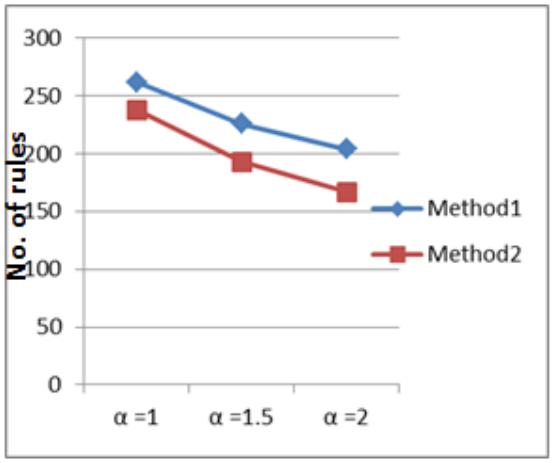


Figure 4: Comparison – No. of rules

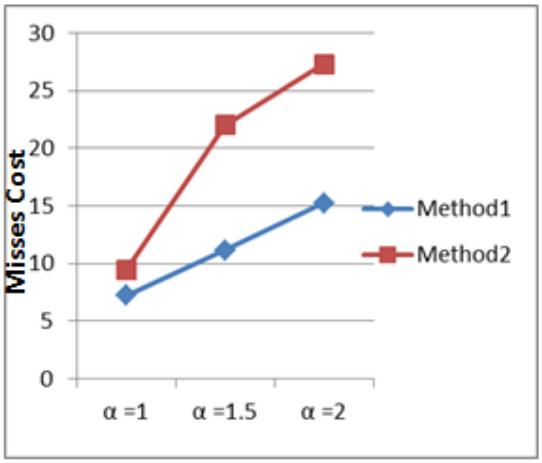


Figure 5: Comparison - MC

Figure 5 shows a comparison of Misses cost for method-1 and mehod-2. The Misses Cost is high in the case of method-2. This can be justified and is acceptable due to the interestingness measure that is considered in the ARIPSO framework, during filtering. The discrimination removal effectiveness of both the methods is nearly identical. This in effect proves that, usage of ontologies in data-mining in general and discrimination prevention in particular is a constructive move, which enhances not only performance, but also the relevance of the mined rules to the context. Similar is Figure 6 which shows the comparison of α- discriminatory rules (direct and indirect) generated out of the two methods. Figure 7 denotes the number of Red-Lining rules generated out of the two methods. From all the analysis performed during the comparison of the two methods, it can be conferred that usage of Ontologies and the additional measures aid to a more efficient filtering of rules, in turn leading to a better discrimination removal methodology.

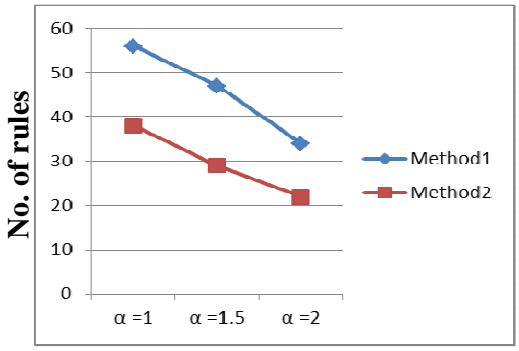


Figure 6: No. of α-discriminatory rules

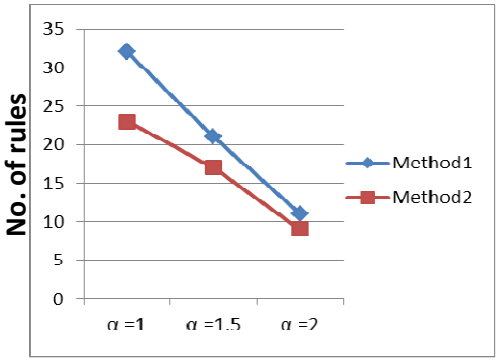


Figure 7: No. of redlining rules

5. CONCLUSION AND FUTURE WORK

Due to the adoption of ARIPSO framework in the proposed system, the interestingness of the rules are preserved, and only those which evidently contribute to the decision making are retained in the resulting set of rules. This proves to be the advantage of the proposed system over the existing discrimination prevention methods. But much remains to be done in this arena to fine tune the proposed method, and some of the enhancements that are noteworthy are-

- Weighted lift and elift measures should be considered instead of flat measures for the attributes of the data set. By doing so, each attribute is assigned a value of importance, which might yield in more efficient method of discrimination prevention.

- Present real case studies for discrimination discovery and prevention using ontologies in data mining.

- Extend the existing approaches and algorithms to a variety of data mining tasks and multiple types of input data. Study and analyse the problem of discrimination prevention in run time, in the case of on-line transaction systems. This calls for attention due to the fact that the discrimination prevention algorithms should cater to the instant of service request and not on a repository of historical data.

- Extend concepts and methods to the analysis of discrimination in social network data. This provides an important case study because of the huge amounts of data that is present in the social networking sites, and their behavioural aspects pertaining to each user.

ACKNOWLEDGEMENTS

We express our gratitude to Dr.Shylaja S S, Professor and Head of Department of ISE, PES Institute of Technology, Bengaluru, India for facilitating during the tenure of the course. We are greatly thankful to all the researchers who have authored numerous literature works which has been a rich plethora of knowledge and forms the basis of this study.

REFERENCES

[1] Sara Hajian, Joseph Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining", IEEE Transactions on Knowledge And Data Engineering, vol. 25, No. 7, July 2013

[2] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining", Proc. 14th ACM Int'l Conf. Knowledge Discovery andData Mining (KDD '08),p. 560, 2008.

[3] Claudia Marinica and Fabrice Guillet,"Knowledge-Based Interactive Postmining of Association Rules Using Ontologies",IEEE Transactions on Knowledge And Data Engineering, vol. 22, No. 6, June 2010

[4] T. Dalenius. The invasion of privacy problem and statistics production: an overview. Statistik Tidskrift, 12:213-225, 1974.

[5] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In Proc. of the 17th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems (PODS 98), Seattle, WA, June 1998, p. 188.

[6] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. Knowledge Information Systems, 33(1): 1-33, 2011.

[7] T. Calders and I. I. Zliobaite, "Why unbiased computational processes can lead to discriminative decision procedures. In Discrimination and Privacy in the Information Society" (eds. B. H. M. Custers, T. Calders, B. W. Schermer, and T. Z. Zarsky), volume 3 of Studies in Applied Philosophy, Epistemology and Rational Ethics, p. 4357. Springer, 2013.

[8] F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling" , Proc. 19th Machine Learning Conf. Belgium and The Netherlands, 2010.

[9] F. Kamiran and T. Calders, "Classification without Discrimination", Proc. IEEE Second Int'l Conf. Computer, Control and Comm. (IC4 '09), 2009.

[10] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," Data Mining and Knowledge Discovery, vol. 21, no. 2, p. 277, 2010.

[11] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring Discrimination in Socially-Sensitive Decision Records," Proc. Ninth SIAM Data Mining Conf. (SDM '09), p. 581, 2009.

[12] D. Pedreschi, S. Ruggieri and F. Turini. Integrating induction and deduction for finding evidence of discrimination. In ICAIL 2009, p. 157. ACM, 2009.

[13] B. Liu, W. Hsu, K. Wang, and S. Chen, "Visually Aided Exploration of Interesting Association Rules," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), p. 380, 1999.

[14] T.R. Gruber, "A Translation Approach to Portable Ontology Specifications," Knowledge Acquisition, vol. 5, p. 199, 1993.

[15] H. Nigro, S.G. Cisaro, and D. Xodo, Data Mining with Ontologies: Implementations, Findings and Frameworks. Idea Group, Inc., 2007.

[16] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering Frequent Closed Itemsets for Association Rules," Proc. Seventh Int'l Conf. Database Theory (ICDT '99), p. 398, 1999.

[17] M. Zaki, "Mining Non-Redundant Association Rules," Data Mining and Knowledge Discovery, vol. 9, p. 223, 2004.

[18] A. Maedche and S. Staab, "Ontology Learning for the Semantic Web," IEEE Intelligent Systems,vol. 16, no. 2, p. 72, Mar. 2001.

[19] D. Pedreschi, S. Ruggieri and F. Turini, "Measuring discrimination in socially sensitive decision records", Proc. of the 9th SIAM Data Mining Conference (SDM 2009), p. 581. SIAM, 2009

[20] P. Samarati, "Protecting respondents' identities in microdata release" IEEE Transactions on Knowledge and Data Engineering, 13(6):1010-1027, 2001.

[21] L. Sweeney. "k-Anonymity:a model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557-570, 2002.

[22] B. C. M. Fung, K. Wang, and P. S. Yu. Top-Down Specialization for Information and Privacy Preservation. In ICDE 2005, p. 205. IEEE, 2005

PLASTICITY OF A GUIDANCE SYSTEM FOR SOFTWARE PROCESS MODELING

Hamid Khemissa[1], Mohamed Ahmed-nacer[2] and Mourad Oussalah[3]

[1, 2] Computer Systems Laboratory, Computer Science Institute,
USTHB University, Bab Ezzouar Algeria
[1]hkhemissa@hotmail.com,
[2]anacer@cerist.dz
[3] Computer Laboratory Nantes Atlantique, Faculty of science Nantes University
Mourad.Oussalah@univ-nantes.fr

ABSTRACT

The need for adaptive guidance systems is now recognized for all process of software development. The new needs generated by the mobility context for software development led these guidance systems to be adapted for. This paper deals with the plasticity of guidance systems or their ability to be adapted to specific development contexts. We propose a Y description for adaptive guidance. This description focuses on three dimensions defined by the material platform, the adaptation form and provided guidance service. Each dimension considers several factors to deduce automatically the appropriate guidance service to a current development context.

KEYWORDS

Plasticity, Adaptation, Development context, Guidance systems plastic.

1. INTRODUCTION AND PROBLEMATIC

The software development organizations are actually confronted to difficulties regarding the development of their applications. Due to technological progress, the developer is considered nowadays as a mobile actor working in various development context using variable platforms. This trend seems interesting from a user perspective, it poses a new problem in software processes engineering. This concern denotes the adaptation ability to the possible variations of the development context. The objective is to support the process by providing software tools to model, improve, assist and automate development activities [1, 2]. For this purpose, the research in the software processes modeling have known a considerable evolution focusing on defining concepts and objectives for modeling and defining Process-Centered Software Engineering Environments [2,3,4]. They agree on the following goals like to facilitate the comprehension and communication process, to describe clearly the roles, responsibilities and interactions between users, to automate the execution of repetitive tasks that do not require the human actor intervention and finally, whatever the support and the development context used, to provide guidance to actors about modeling and handling a software process. According to the aim and orientation given to the software process, it is possible that other concepts such as strategy, organization and guidance can be described in the software process meta-model.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 49–63, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4905

For this, it is necessary to assist developers and to ensure plasticity of the guidance systems [5, 6] by their ability to adapt to the current development context in respect of their usefulness. Also, usefulness is not limited to performance criteria in the tasks accomplishment, it relates rigorously to satisfaction services offered to developers. By development context, we mean the triplet (material platform, developer profile, activity context). Usefulness refers to the ability of a guidance system to allow the developer to reach his objective preserving consistency and product quality in software development.

In this perspective, a rigorous guidance system targets two basic aspects: 1) The progress control of the software process development regarding the temporal constraints of the activity and the consistency of the results, and 2) the guidance interventions adapted to the specific needs within the development context in progress.

Section 2 of this paper presents a synthesis of similar work and describes the current tend. Section 3 presents our approach of the Y adaptive guidance modeling while section 4 describes the implementation process of the adaptive guidance. Section 5 describes the Plasticity of Guidance Meta model (PGM) and section 6 presents the practical cases study of the adaptive guidance. It ends with a conclusion and future prospects.

2. RELATED WORKS AND CURRENT TEND

Several process-centered environments [7, 8, 9] deal with the assistance aspect in the support of the software product development. However, the provided guidance is not often adapted to the development context profile. The orientations of the guidance are defined on the basis that the human actor, regardless of his profile (qualifications and behavior), has a central role in the progress of the development process.

Among this new generation of the software process engineering, we can invoke the following meta-models and modeling environments: SPEM [10] and APEL [8] considered as the most representative in the software process modeling, RHODES [7][11] that uses basic concepts closest to those introduced by the proposed approach.

SPEM meta-model introduced the concept of "Guidance". According to SPEM, the guidance is a describable element which provides additional information to define the describable elements of modeling. However, the proposed guidance is not suitable to the development context's profile (role, qualifications and behavior). The guidance is rather defined in an intuitive way. ADELE/APEL proposes a global assistance of proscriptive type without considering the development context profile and automates part of the development process using triggers. RHODES/PBOOL+ uses an explicit description of a development process. The activities are associated to a guidance system with various scenarios of possible realization.

An effective support to software process depends on several factors, in particular the personalization and adaptation factor. The definition of a process with an active guidance for automation and coherency control would be effective if it can be adapted to each development context. The platform, tasks and developers characteristics may considerably vary. An improved productivity and development process adaptation would be possible, if a process can be adapted considering the fact that these characteristics can be exploited.

Actually, there are Process Centered Software Engineering Environments (PSEE) allowing changes during the execution, where the developer is in a position to predict the execution model before running it. However, these models do not provide appropriate performance models. Some PSEEs use a guidance description structured in phases like prescribing systems or proactive

systems to control the operations carried out by the developer. Nevertheless, they are essentially limited to the adaptive guidance aspect to current development context.

Taking into account specific criteria for an adaptive guidance, we have classified these limits through several criteria describing explicitly the basic concepts linked to the adaptive guidance [5, 6, 12]. To realize the effectiveness of plasticity concept of the guidance system supported by its adaptation ability to current development context, we refer to the studied meta-models and modeling environments [12, 13]. The selected criteria are defined by:

- Global guidance core: The basic guidance is defined as a global orientations core regardless the profile of both the activity and the actor.
- Developer profile oriented guidance: the guidance orientations are defined on the basis that the human actor, regardless his profile, has a central role in the progress of the development process.
- Context development guidance: The selection of the appropriate type of guidance is more often not adapted nor suitable to a current context.
- Guidance types: the selection of guidance types remains defined in a manual and in an intuitive way. It depends on the experience and on the informal personality of the project manager.
- Plasticity of guidance: the guidance functions are defined and offered on the basis that the human actor always operates in a uniform development context.

To respond to these limits, one currently tries to offer more flexibility in the language of software process modeling. This tendency results in the idea to define interventions of direct and adaptive assistance in particular contexts during the progress of software process. In considering the principal limitations of PSEEs and essential characteristics of our approach in particularly the context adaptation aspect, a comparative table of the studied meta-models is a follows.

Table 1. Comparative table of the studied meta-models.

Meta model / Criteria	ADELE/APEL	RHODES / PBOOL+	SPEM
Global guidance core	Global	Global	Global
Developer profile oriented guidance	Not adapted	Considered strategy Model	Not adapted
Context development guidance	Not adapted	Adapted	Not adapted
Guidance types	Not invoked	Associated with a specific guide system	Intuitive selection
Plasticity of guidance	Not covered (Single Platform)	Not covered (Single Platform)	Not covered (Single Platform)

The current tendency is that developers would like to have integrated environments that are suitable to specific needs according to the characteristics of the development context. However, despite the necessity imposed by technological evolution, the provided efforts to develop such environments remain an insufficient contribution. This generation of guidance environment still interests researchers in defining new concepts and objectives of the software process modeling [4, 14, 15].

Our work proposes an approach to define adaptive guidance modeling in software process. The proposed approach concepts are described through a meta-model denoted PGM (***Plasticity of Guidance Meta model***). The information provided must be adapted to the development context profile. They must guide the developer during the software process development through suitable actions and decisions to undertake with corrective, constructive or automatic intervention [12]. Its adaptation is explicitly described by three plasticity dimensions defined by the development context, the adaptation form and the provided service.

3. THE ADAPTIVE GUIDANCE IN Y

A guidance system may be processed in many different ways according to the perspective guidance to provide interveners with development context. Thus, there are generally several possible assistance models, each of them with a particular relevance and need. This vision denotes the plasticity of guidance system, and its ability to adapt to their development context. The plasticity concept describes its capacity to adapt to the intrinsic variations of required conditions in terms of usefulness [16, 17].

In this context, we propose a description in Y of the adaptive guidance. This description will focus on the three considered dimensions. Each dimension considers several factors to deduce automatically the appropriate guidance service according to the current context. It is schematically described as follows:

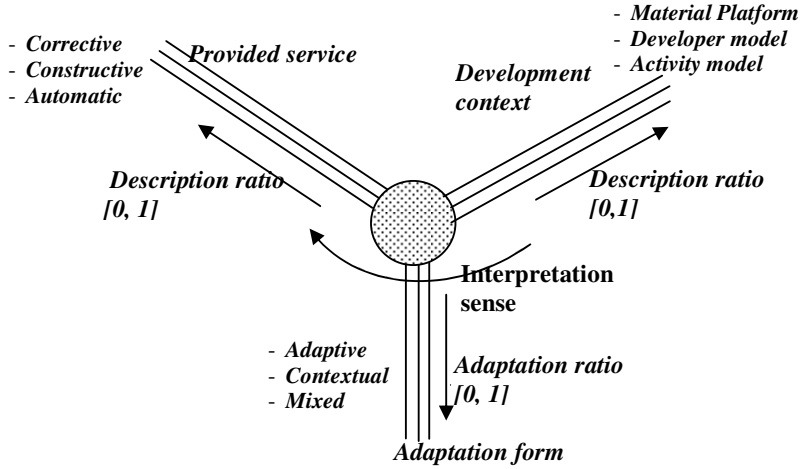


Figure 1. Adaptive guidance description in Y.

The principle of our approach is to generate, from the development context related to the specific data for each defined models according to the retained adaptation form, the guidance interventions (corrective, constructive, automatic) adapted to the current development context.

3.1. The basic conceptual model description

The conceptual model highlights the guidance plasticity aspect through the adaptation form described by the inherent relationship between the three considered dimensions.

3.1.1. The development context

A guidance intervention is provided according to an object or set of objects. An object is associated to:

- The hardware and software platform: described by the computing resources, the software services and the interaction and communication modes.

- The activity model: models the structure and the workflow, they are defined by a progression mode in the activity ensuring that all tasks can be performed under control in a preset order established by the designer and a temporal progression mode specifying deadlines for completion.

- The developer model: defines the specific properties of each developer. These properties can be either static or dynamic. The static aspect refers to the user characteristic as his role, his business competence and his familiarity with the software process. The dynamic aspect refers to the behavior of using the guidance system, by the fact to execute, to define or to complete the software process resource and the user's reaction to a guidance message.

The description performance rate of these factors is evaluated by considering each identified object as concepts, principles, procedures, and resources. These guidance objects represent the basis of different guidance interventions related to a particular situation. This performance serves as the selection of the adaptation form to retain and guidance service to provide to the user.

3.1.2. The adaptation form

Each guidance intervention is done according to the retained adaptation form. It relates to a specific situation described by the development context description. Our modeling approach allows the following guidance adaptation forms:

- **Contextual guidance**: intervention is provided dynamically according to the material platform and activity models and the state of the process. The adaptation rate is related to the model description rate of the activity and the material platform. The guidance intervention doesn't consider the developer model (e.g.: to avoid inconsistency during the affectation of a resource).

- **Adaptive guidance**: intervention is provided according to the developer model and the material platform specificity (e.g.: the user asks for explanations on his choice). The adaptation rate is related to the developer and material platform models description.

- **Mixed guidance**: intervention is provided according to the development context (e.g.: to guide the developer on the sequencing principle during the software process progression). This form describes the highest adaptation rate. This rate is evaluated on the basis of the developer, activity and material platform models description.

The adaptation form performance is described by a strong coupling between the development environment and guidance system. It determines the relevance and precision of the guidance provided to developers.

This criterion is directly related with the adaptive guidance system concept. Through a strong coupling, the system would deduce the guidance context and can therefore extract useful and helpful information to the user.

3.1.3. The provided service

The guidance system offers several service types in relation to a defined context by the current development and adaptive form. The provided services are corrective, constructive or automatic order.

- **Control and taking corrective initiative:** protect the user of his own initiatives when they are inadequate under progress.

- **Control and taking constructive initiative:** the ability to take positive initiatives, executing and combining the performance of operations without the user intervention.

The guidance adaptation performance associated to a development environment is done by enrichment or reduction of the possible offers of the guidance. Among these offers, we have:

- **The directive guidance:** to show the developer how to execute a task by an adaptive control of the guidance system, specifying the steps of an activity or the whole process development.

- **Retroaction**, to offer the developer more information on the activities context (e.g.: new available resources) or on the progress state of his work (progression of an activity).

- **Explanation**, to offer explanations about a guidance object at the request developer. (e.g.: the activities coordination of the software process).

- **Reminding**, to remind the developer some principles or procedures on the sequencing of the activities or their activation conditions when the system detects a conflict or inconsistency.

- **Automatic guidance:** analyze the impact projection to define the solution to consider in order to avoid deadlocks or delays, by the fact to start, suspend, discontinue or continue ongoing actions to avoid conflict.

These services can be combined. They may be temporary, permanent or left under the developer control.

The usefulness rate is evaluated by the degree of the performance description of the development context and adaptation form.

4. THE ADAPTIVE GUIDANCE FUNCTIONING

The implementation of the proposed adaptive guidance is done according to the interpretation sense (see Figure 1). The selection of the adaptation form is relatively based on the description rate of the development context elements. The provided service will focus on the retained adaptation form, relatively to the concepts' interpretation related to the development context. The adaptation of guidance system to the development context can only concern a subset of the latter. In this case, we will talk about a reduced service associated to the provided guidance linked to the current context.

In all cases, the operating strategy in the adaptation is done by the reduction or enrichment of the provided guidance service. Intuitively, we consider the three following adaptation strategies:

- Service plus (or enrichment): is a strategy to enhance the offered guidance services to support an expressed description of the development context.

- Service minus (or reduction): is to remove guidance items due to a limited or non-critical description of the context.

- Service poly: generates several possible forms of service. This strategy is supported by the performance rate principle of the provided services.

These Strategies are accomplished according to the political autonomy given to the guidance system in respect of the context conditions and the developer choice. The choice of this policy is made with regard to the performance criteria of the three guidance dimensions. The performance degree of each component varies from 0 to 1. The implementation of this policy is based on an adaptation mode expressed by a set of rules of the ECA form (Event, Conditions, Actions). For each Event, if required Conditions then propose Actions.

4.1. The instantiation process

The instantiation process of the proposed generic model will focus on concrete specific use cases. It described how to generate the guidance service in adequacy with the current development context through concrete situations.

Situation 1: to support the progression of high performance developer which evolves on an average order platform and takes in charge a simple activity. Therefore, these two factors are practically without effect, the adaptation of the provided service, based on the developer performance choice, reaches an adaptive guidance of a corrective order (see Figure 2).

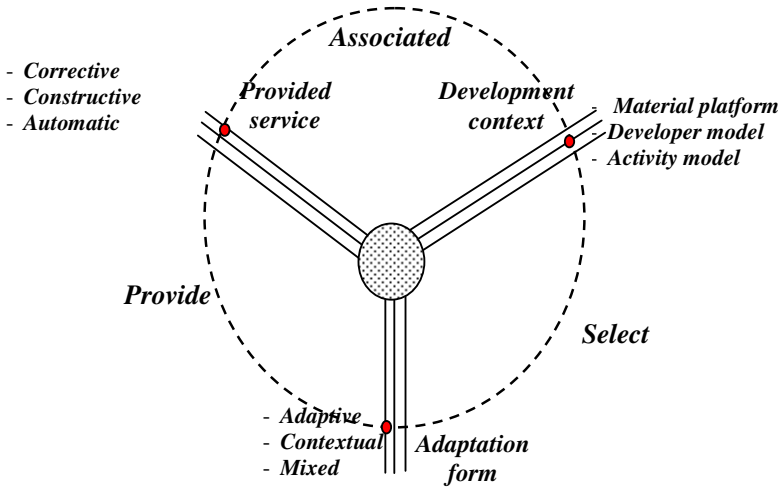


Figure 2. Description of the adaptation scenario : situation 1.

Situation 2: Assuming that the development context migrates toward a development platform relatively limited and that the two other factors are always of an average order and practically without effect. The adaptation of the provided service will rather be on a contextual form associated to corrective as well as constructive guidance (see Figure 3).

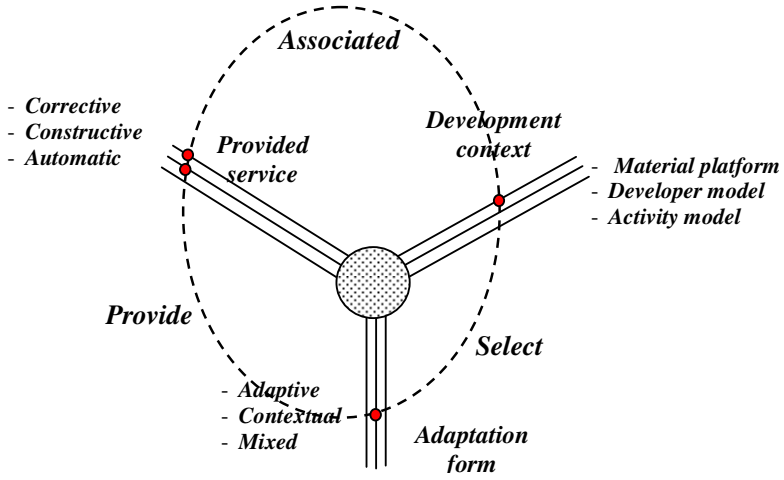


Figure 3. Description of the adaptation scenario : situation 2.

5. PLASTICITY OF GUIDANCE META MODEL

Our modeling approach PGM (Plasticity of Guidance Meta model) is defined with reference to the identified limitations of studied PSEEs. The essential characteristic of our approach is to consider the plasticity principle in the development context, defined by the description of its three models [15, 17].

In this context, our meta-model is based on the conceptual model of a software process enriched by the plasticity of the adaptive guidance element. It controls the smooth running of the activities and provides adaptive guidance to the development context.

The Adaptive guidance management addresses the three defined dimensions by the development context, adaptation form and provided guidance service. Each dimension considers several factors to deduce automatically the appropriate guidance service to the current context. It has an operating strategy supported by three services.

The first service 'Service Plus' role is to enhance the guidance function to support the current situation. The second service 'Service Minus' is to adapt by reducing the guidance function to a particular context. Finally, the third service 'Service Poly' generates, according to the current context, the most suitable form of the offered guidance function.

The guidance strategy evolves according to the political autonomy given to the guidance system respecting the application conditions. The implementation of this policy is based on an adaptation mode expressed by a set of rules of ECA form (Event, Conditions, Actions). For each in the execution context, if required conditions related to the context and the adaptation form then launch guidance strategy to generate the most appropriate service.

The proposed meta-model aims to generate the adapted guidance interventions to the development context in relation to the considered properties and specific data for each defined model (see Figure 4).

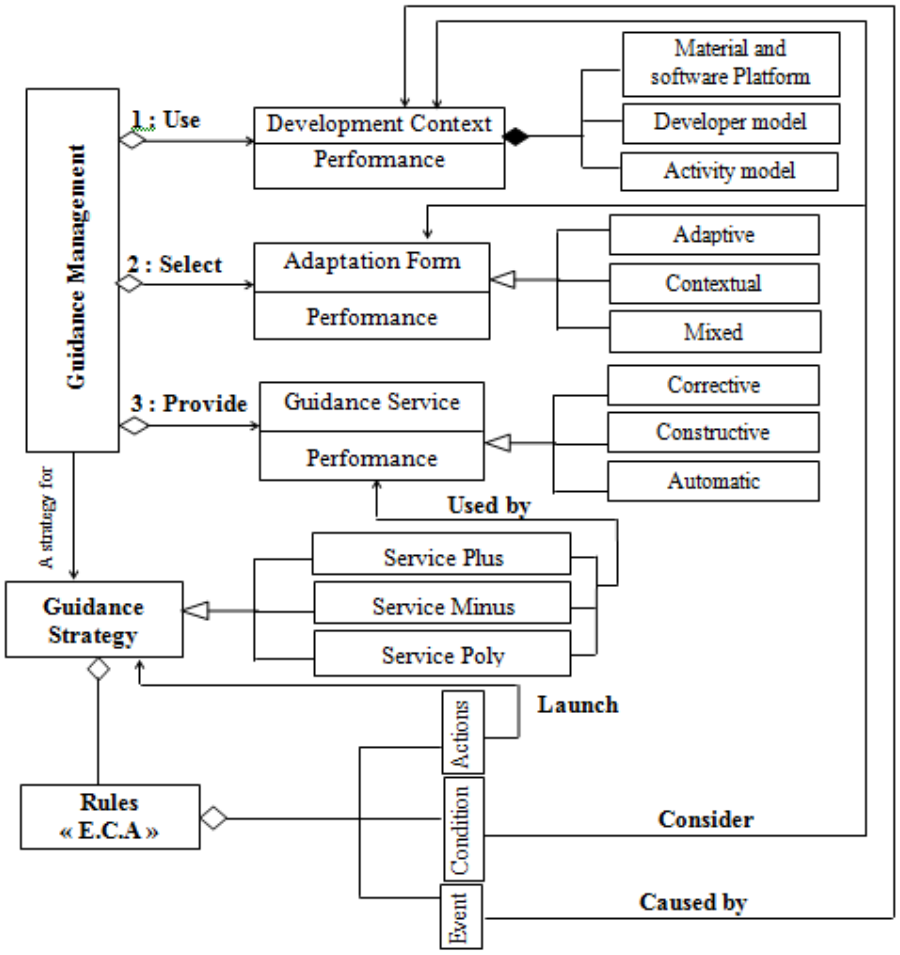


Figure 4. Plasticity of Guidance Meta model.

6. THE PRACTICAL INTERPRETATION

Considering the software process model "Activity test", the process "Activity test" in the software development is composed of several types of tests such as: Integration test and Unitary test. Each receives as input a test plan and provides a test report. For each type of test, there is a manager, responsible of the execution.

The activity process "Activity test" is described by a performing tree given in Figure 5. We notice that the activity test starts the execution of subactivities "Unitary test" then "Integration test". The unitary test launches in parallel the execution of tasks "Test unit".

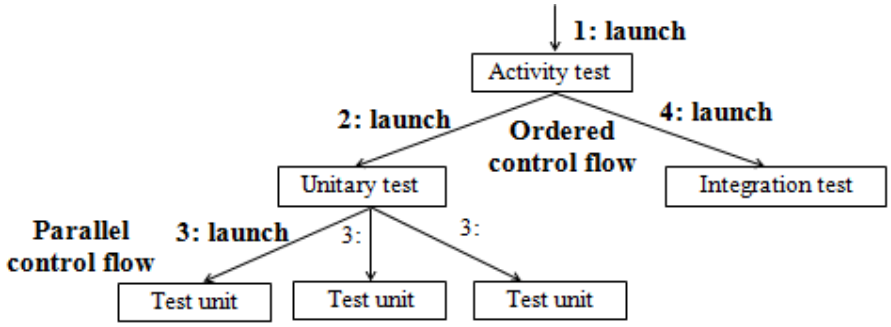


Figure 5. The activity test process.

To simplify our example, we consider the execution process of the unitary component test. The application of the activity “Unitary test", requires the list of components. It calls the tool that will create the necessary environment to carry out the actual execution of the “Unitary test", as the state diagram, the test variables, etc. ... the activity "Unitary test" launches in parallel the different tasks "Test unit" where an event signals the beginning of the “Test unit” execution. Finally, the ended event is broadcast.

The adaptive execution process of the activity "Unitary test", regarding our adaptive guidance approach is described by Figure 6.

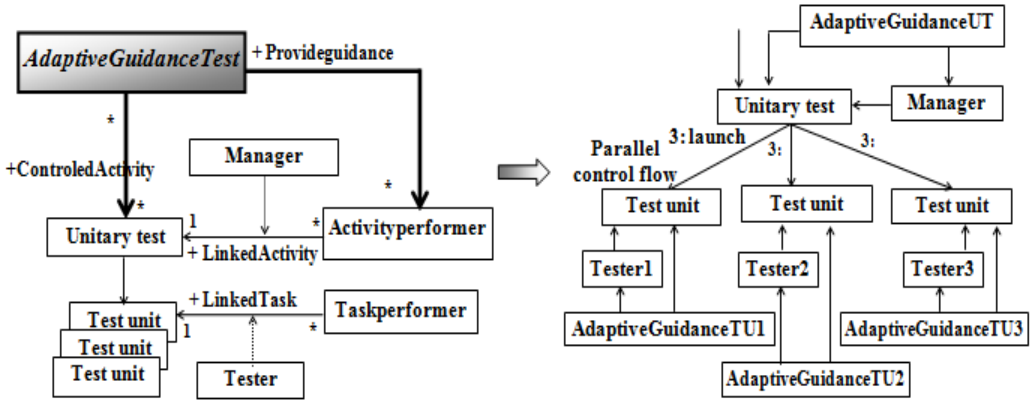


Figure 6. The adaptive execution process “ Unitary test”.

The adaptive guidance is linked to the manager or to each tester according to the current development context profile defined by its material platform, the activity model and developer model. We explain this adaptive approach through the following situation; the testers have the same role “test unit” with identical activity model. However, the developer’s qualification and the material platform specificity differ from one development context to another. According to the current context, it happens to enrich or reduce the appropriate guidance intervention or generate several possible forms of guidance services.

We consider three situations with tester’s qualification defined respectively as high, medium, and low. The study case is related to launch the test unit without having all the input data, by selecting the appropriate test variables and generating the unit test report. The adaptive guidance process related to each qualification case is described as follows:

1. For a development context with high qualification tester and a high material platform performance: the tester starts the test unit process on the basis of the defined plan by taking his proper initiatives. The development context evaluation allows deducing the adaptation form to retain and the guidance service to provide. In this case, we adopt the adaptation guidance form and the provided guidance intervention is thus of a corrective order. The corrective intervention is provided to inform the manager of the setback and remind him of the corresponding unitary test diagram. The manager remains free to take into account the intervention.

2. For a development context with an average skill tester and an acceptable material platform performance: the tester starts the test unit process by applying rigorously the defined test plan. The evaluation of such context results in a contextual guidance form and the provided guidance intervention is thus of a constructive order. The guidance system analyzes the current context of the task, evaluates the impact and consequence of the delay caused in comparison with possible margins and offers a possible solution to the manager (solution: the guidance proposes to cancel the launch of the current test unit and generates a new execution plan according to the rate of delay and possible margins). The construction solution is not definite; it should be validated by the manager.

3. For a development context with a low qualification tester and a reduced material platform performance: the tester starts the test unit process by applying reliably the defined test plan. The development context evaluation results in a mixed guidance adaption form and the provided guidance intervention is thus of an automatic order. The guidance system analyzes the current context, cancels the launch of the “test unit” task, evaluates the impact and consequence of the delay caused in comparison with the possible margins and automatically updates the execution plan of "unitary test" activity.

6.1. The digital application

The practical definition of the adaptive guidance type for each considered profile is deduced by a quantitative process of the characteristics in relation to the basic models (materiel platform, activity model, developer model). The considered example is processed as follows.

Each profile is semantically described in table (see Table 2). The semantics evaluation and the weighting are determined by the project manager under the specification of an ongoing project [14]. To scan the semantics evaluation, we associate the weighting related to the interest granted to each attribute.

Table 2. The profiles evaluation.

Development context	Features	Context Profile 1	Context Profile 2	Context Profile 3	Context Profile 4	Context Profile 5	W[i]
Material Platform	Development System Constraint	Low	Medium	High	High	Low	**P2**
	Software Tools	Low	Medium	Low	Provided	Provided	**P1**
	Memory Constraint	High	Medium	Medium	High	Medium	**P3**
Developer	Role	No effect	Classic	Critique	Critique	No effect	**P4**
	Competence	High	Medium	Low	Low	High	**P1**

Model	Familiarity with Software Process	Quite Acceptable	Medium	Low	Low	Acceptable	**P1**
	Behavior for assistance	Most Appropriate	Satisfying	Inadequate	Adequate	Inadequate	**P2**
Activity Model	Density of tasks in the activity	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable	**P3**
	Complexity Level	Medium	Medium	Medium	Medium	Medium	**P2**

With $W[i] \in [1, 5]$. Where Pi represents the computing value.

Considering the similar principle such as the COCOMO model, the quantification of each profile's feature is on the data range] 0, 2 [, (see Table 3). This quantification is usually based on the impact of each feature.

It is usually done through three levels, described by high, medium or low contribution, applying the following rules:

1: middle order impact / <1: positive impact / >1: negative impact.

Table 3. The profiles quantification.

Development context	Features	Context Profile 1	Context Profile 2	Context Profile 3	Context Profile 4	Context Profile 5	W[i]
Material Platform	Development System Constraint	**0.75**	**1.00**	**1.30**	**1.30**	**0.75**	**1**
	Software Tools	**1.20**	**1.00**	**1.25**	**0.80**	**0.80**	**3**
	Memory Constraint	**1.40**	**1.00**	**1.00**	**1.60**	**1.00**	**2**
Developer Model	Role	**0.40**	**1.00**	**1.90**	**1.70**	**0.40**	**4**
	Competence	**0.20**	**1.00**	**1.70**	**1.70**	**0.25**	**3**
	Familiarity with Software Process	**0.40**	**1.00**	**1.60**	**1.60**	**0.30**	**3**
	Behavior for assistance	**0.20**	**0.80**	**1.70**	**0.75**	**1.60**	**1**
Activity Model	Density of tasks in the activity	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	**2**
	Complexity Level	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1**

In this stage of profiles' process, and in case of simple profiles' samples, we can proceed to associate each considered development context profile to the appropriate guidance adaptation form and guidance service.

The guidance profile (GP) associated to each profile class is based on the following formula:
$$GP(Px) = \Sigma Ai\ Wi / 2*\Sigma Wi \text{ avec } i=1 \text{ to } n$$

With :

Ai: the feature value.
Wi: the associated weighting.
Px : the associated profile.

The adaptation form and the guidance profile of each considered development context profile based on the evaluation of each model and GP value is given by (see Table 4).

Table 4. The associate guidance profile.

	Context Profile 1	**Context Profile 2**	**Context Profile 3**	**Context Profile 4**	**Context Profile5**
Associated Adaptation form	Adaptive	Contextual	Mixed	Mixed	Adaptive
Guidance Profile (GP)	0.333	0.485	0.721	0.673	0.315
Associated guidance profile	Corrective	Constructive	Automatic	Automatic	Corrective

It should be noted that the value of GP ranged from 0 to 1 and the range associated with each type of guidance is defined by the fixed limits to each guidance type. If the range of corrective guidance is fixed between 0 and 0.35 and the range of the constructive guidance is between 0.36 and 0.65, we automatically associate a corrective guidance to profile P1 and P5, and a constructive guidance to profile P2 and automatic guidance to profile P3and P4.

However, in case of a very important population, and for the aim of optimizing profile classes, it is recommended to proceed in the gathering and classification of the provided development profile and reasoning in relation to generated classes.

7. CONCLUSION

Our main purpose in this article is to propose a plasticity of a guidance system for software process modeling. This plasticity is highlighted through a description in Y of our adaptive guidance. This description will focus on three dimensions defined by the material platform, the adaptation form and the provided service. Each dimension considers several factors to deduce automatically the appropriate guidance service according to the current context. The proposed approach concepts are described through a meta-model denoted PGM (***Plasticity of Guidance Meta model***). The proposed meta-model aims to generate the adapted guidance interventions to the development context in relation to the considered properties and specific data for each defined model.

The operating strategy in the adaptation is done by the reduction or enrichment of the provided guidance service. Intuitively, we consider the three adaptation strategies (Service Plus, Service Minus, and Service Poly). The guidance strategy evolves according to the political autonomy given to the guidance system respecting the application conditions. The implementation of this policy is based on an adaptation mode expressed by a set of rules of ECA form.

A perspective to this work concerns, at first, the necessity to estimate the productivity and cost due to the adaptation of guidance system.

In a second step, we will ensure the development of semantic rules which allow swapping through different guidance profiles, either statically by adjustment of guidance parameters or dynamically through the performer behavior.

REFERENCES

[1] Ivan Garcia and Carla Pacheco « Toward Automated Support for Software Process Improvement Initiatives in Small and Medium Size Enterprises ». Book chapter. Software Engineering Research, Management and Applications 2009 Volume 253/2009, pp. 51–58. c_ Springer-Verlag Berlin Heidelberg 2009. ISBN: 978-3-642-05440-2.

[2] Kirk, D.C, MacDonell, S.G., & Tempero, E. 2009 Modeling software processes - a focus on objectives, in Proceedings of the Onward, 2009. Conference. Orlando FL, USA, ACM Press, pp.941-948.

[3] Benoît COMBEMALE, Xavier CRÉGUT, Alain CAPLAIN et Bernard COULETTE. Towards a rigorous process modeling with spem. Dans ICEIS (3), pages 530–533, 2006

[4] Hamid Khemissa, Mohamed Ahmed-Nacer, Mourad Daoudi, 2008. A Generic assistance system of software process. In International Conference on Software Engineering: Software Engineering. SE 2008, Feb 12-14-2008, Innsbruck, Austria.

[5] Calvary, G., Coutaz, J., Thevenin, D., Limbourg, Q., Souchon, N., Bouillon, L., Florins, M., Vanderdonckt, J.: Plasticity of User Interfaces: A Revised Reference Framework. In: TAMODIA 2002 (2002).

[6] Joëlle Coutaz, EICS '10. User interface plasticity: model driven engineering to the limit!. Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing systems. June 2010.

[7] Coulette B., Crégut X., Dong T. B. T. and Tran D. T., "RHODES, a Process Component Centered Software Engineering Environment", ICEIS2000, 2nd International Conference on Enterprise Information Systems, Stafford, pp 253-260, July 2000.

[8] Jacky Estublier, Jorge Villalobos, Tuyet Le Anh, Sonia Jamal-Sanlaville and German Vega. An Approach and Framework for Extensible Process Support System. In Proceedings 9th European Workshop on Software Process Technology (EWSPT 2003), Helsinki, Finland, 2003-09-01.

[9] Hans-Ulrich Kobialka, « Supporting the Software Process in A Process-centered Software Engineering environment », Upgrade-cepis.org/issues/2004/5/upgrade-v VOL; V n° 5 October 2004.

[10] OMG. Inc. Software and System Process Engineering Meta-Model Specification version 2.0: Formal/2008-04-01.

[11] Tran Hanh Nhi, Bernard Coulette, Xavier Crégut, Thuy Dong Thi Bich, Thu Tran Dan. Modélisation du méta-procédé RHODES avec SPEM. Dans : Recherche Informatique Vietnam-Francophone (RIVF'03), Hanoi, Vietnam, 2003.

[12] Hamid khemissa, Mohamed ahmed nacer & Mourad Oussalah «Adaptive Guidance System for SPEM ». The First International Conférence on Information Technology Convergence and Services; ITCS, SIP, JSE 2012 pp. 429-441, Bangalore, India.

[13] Hamid Khemissa, Mohamed Ahmed-Nacer, Mourad Oussalah «Adaptive Guidance based on Context Profile for Software Process Modeling». Information Technology and Computer Science, , July 2012 in MECS 2012.

[14] Grambow, Gregor and Oberhauser, Roy and Reichert, Manfred (2011) Enabling Automatic Process-aware Collaboration Support in Software Engineering Projects. In: Selected Papers of the ICSOFT'11 Conference. Communications in Computer and Information Science(CCIS).

[15] Clarke, Paul and O'Connor, Rory (2011) An approach to evaluating software process adaptation. In: 11th International SPICE Conference on Process Improvement and Capability dEtermination, 30 May - 1 jun 2011, Dublin, Ireland. ISBN 978-3-642-21233-8.

[16] Sottet, J.-S., Calvary, G., Coutaz, J., Favre, J.-M. A Model-Driven Engineering Approach for the Usability of User Interfaces. In Proc. Engineering Interactive Systems (EIS2007), J. Gulliksen et al. (eds), LNCS 4940, (2007), 140-157

[17] Ferry, N. Hourdin, G., Lavirotte, S., Rey, G., Tigli, J.- Y., Riveill, M. Models at Runtime: Service for Device Composition and Adaptation. In 4th International Workshop Models@run.time, Models 2009(MRT09).

AUTHORS

Hamid Khemissa is a full associate professor at Computer Systems Department, Faculty of Electronics and Computer Science, USTHB University, Algiers. He is member of the software engineering team at computer system laboratory LSI, USTHB. His current research interests include Software Process Modeling and Software Modeling Assistance.

Mourad Chabane Oussalah is a full Professor of Computer Science at the University of Nantes and the chief of the software architecture modeling Team. His research concerns software architecture, object architecture and their evolution. He worked on several European projects (Esprit, Ist, ...). He is (and was) the leader of national project (France Telecom, Bouygues telecom, Aker-Yard-STX, ...). He earned a BS degree in Mathematics in 1983, and Habilitation thesis from the University of Montpellier in 1992.

Mohamed Ahmes-Nacer is a full Professor at USTHB (Algiers's University). He is in charge of the software engineering team. He published extensively and his current research interests include process modeling, information systems, software architecture based components and service web development.

INTENTIONAL BLANK

CONTENT BASED IMAGE RETRIEVAL : A REVIEW

Shereena V.B.[1]and Julie M. David[2]

[1,2]Asst.Professor, Dept of Computer Applications,
MES College, Marampally, Aluva, Cochin, India
[1]shereenavb@gmail.com
[2]julieeldhosem@yahoo.com

ABSTRACT

In a content-based image retrieval system (CBIR), the main issue is to extract the image features that effectively represent the image contents in a database. Such an extraction requires a detailed evaluation of retrieval performance of image features. This paper presents a review of fundamental aspects of content based image retrieval including feature extraction of color and texture features. Commonly used color features including color moments, color histogram and color correlogram and Gabor texture are compared. The paper reviews the increase in efficiency of image retrieval when the color and texture features are combined. The similarity measures based on which matches are made and images are retrieved are also discussed. The paper discusses effective indexing and fast searching of images based on visual features.

KEYWORDS

CBIR, Color moments, Color histogram, Color correlogram, Gabor filter, Precision, Recall.

1. INTRODUCTION

Image Processing involves changing the nature of an image in order to improve its pictorial information for human interpretation and render it more suitable for autonomous machine perception [1].The advantage of image processing machines over humans is that they cover almost the entire electromagnetic spectrum, ranging from gamma to radio waves where as human eye is limited to the visual band of the electromagnetic spectrum. They can operate on images generated by sources like ultrasound, electron microscopy, and computer-generated images. Thus image processing has an enormous range of applications and almost every area of science and technology such as medicine, space program, agriculture, industry and law enforcement make use of these methods. One of the key issues with any kind of image processing is image retrieval which is the need to extract useful information from the raw data such as recognizing the presence of particular color or textures before any kind of reasoning about the image's contents is possible.

Early work on image retrieval can be traced back to the late 1970s. In 1979, a conference on Database Techniques for Pictorial Applications was held in Florence[2]. Early techniques were not generally based on visual features but on the textual annotation of images, where traditional database techniques are used to manage images. Many difficulties were faced by text based retrieval, since volume of digital images available to users increased dramatically. The efficient management of the rapidly expanding visual information became an urgent problem. This need

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 65–77, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4906

formed the driving force behind the emergence of content-based image retrieval techniques (CBIR).

CBIR is a technique which uses visual contents to search images from an image database. In CBIR, visual features such as colour and texture are extracted to characterise images. CBIR draws many of its methods from the field of image processing and computer vision, and is regarded as a subset of that field. In CBIR, visual contents are extracted and described by multidimensional feature vectors. To retrieve images, users provide the retrieval system with example images. The system changes them into internal representation of feature vectors. The similarities or differences between feature vectors of the query examples and those of the images in the database are calculated and retrieval performed with an indexing scheme .The indexing scheme is an efficient way to search for image database. Recent retrieval systems have incorporated user's relevance feedback to modify the retrieval process.

The tasks performed by CBIR can be classified into pre-processing and feature extraction stages. In Pre-processing stage, removal of noise and enhancement of some object features which are relevant to understanding the image is performed. Image segmentation is also performed to separate objects from the image background. In Feature Extraction stage, features such as shape, colour, texture etc. are used to describe the content of the image. This feature is generated to accurately represent the image in the database. The colour aspect can be achieved by the techniques like moments, histograms and correlograms. The texture aspect can be achieved by using transforms or vector quantization. Similarity Measurement is also done in this stage. ie. the distance between query image and different images in the database is calculated and the one with the shorter distance is selected [3]. Similarity measurement can be formulated as follows.

Let { F(x,y):x=1,2,…X, y=1,2,….Y}be a 2D image pixel array.

For colorimages , F(x,y) denotes the color value at pixel (x,y)

ie,{F(x,y)=FR(x,y),FG(x,y), FB(x,y)}

For black and white images F(x,y) denotes the gray scale intensity at (x,y).

The problem of image retrieval can be quoted mathematically as follows:

For a query image Q, we find an image T from the image database such that the distance between corresponding feature vectors is less than the specified threshold t.

ie,D(Feature(Q),Feature(T))<=t

There is a lot of research being done in the field of CBIR in order to generate better methodologies for feature extraction. In this paper, a study of different color and texture descriptors for content-based image retrieval is carried out to find out whether a combination of different features gives better results.

The rest of this paper is organized as follows. In Section 2, we discuss previous work in CBIR. In Section 3, we explain feature extraction and representation methods. Section 4 explains combination of features, Section 5 explains Performance evaluation and indexing schemes and finally, conclusions are given in Section 6.

2. LITERATURE REVIEW

Researchers have proposed different methods to improve the system of content based image retrieval. Ryszard S. Chora´s[3]stated in his paper that the similarity of the feature vectors of the query and database images is measured to retrieve the image. M. Stricker, and M. Orengo, have shown that[4] the first order (mean), the second (variance) and the third order (skewness) color moments have been proved to be efficient and effective in representing color distributions of images. In his paper J. Huang, et al., [5] proposed the color correlogram to characterize not only the color distributions of pixels, but also the spatial correlation of pairs of colors. Deepak S. Shete1, Dr. M.S. Chavan[6] proposed that the ability to match on texture similarity can often be useful in distinguishing between areas of images with similar color (such as sky and sea, or leaves and grass). Fazal Malik, Baharum Baharudin[7] proposed a CBIR method which is based on the performance analysis of various distance metrics using the quantized histogram statistical texture features. The similarity measurement is performed by using seven distance metrics. The experimental results are analysed on the basis of seven distance metrics separately using different quantized histogram bins such that the Euclidean distance has better efficiency in computation and effective retrieval. This distance metric is most commonly used for similarity measurement in image retrieval because of its efficiency and effectiveness.

In the paper of Manimala Singhaand K.Hemachandran [8], they presented a novel approach for Content Based Image Retrieval by combining the color and texture features called Wavelet-Based Color Histogram Image Retrieval (WBCHIR). Similarity between the images is ascertained by means of a distance function. The experimental result shows that the proposed method outperforms the other retrieval methods in terms of Average Precision. Md. Iqbal Hasan Sarker and Md. ShahedIqbal [9] proposed that using only a single feature for image retrieval may be inefficient. They used color moments and texture features and their experiment results demonstrated that the proposed method has higher retrieval accuracy than the other methods based on single feature extraction. N.R. Janani and Sebhakumar P. suggests [10] a content-based image retrieval method which combines color and texture features in order to improve the discriminating power of color indexing techniques and also a minimal amount of spatial information is encoded in the color index. The motivation behind this paper is a study on the works done by early researchers in the field of content based image retrieval based on color and texture features.

3. FEATURE EXTRACTION AND REPRESENTATION

Features are properties of images such as colour, texture, shape, edge information extracted with image processing algorithms. A single feature does not give accurate results, but a combination of features is minimally needed to get accurate retrieval results.

3.1 Color

The most widely used visual feature in image retrieval is color feature. Color feature is relatively robust to background complications. Each pixel can be represented as a point in 3D color space. Commonly used color space include RGB, CIELab where “L” value for each scale indicates the level of light or dark, “a” value redness or greenness, and “b” value yellowness or blueness, HSV (Hue, Saturation, Value).

In the RGB color space, a color is represented by a triplet (R,G,B), where R gives the intensity of the red component, G gives the intensity of the green component and B gives the intensity of the blue component. The CIE Lab spaces are device independent and considered to be perceptually

uniform. They consist of a luminance or lightness component (L) and two chromatic components a and b or u and v. HSV (or HSL, or HSB) space is widely used in computer graphics and is a more intuitive way of describing color. The three color components are hue, saturation(lightness) and value (brightness). HSV colour model describes colours in terms of their shades and brightness (Luminance). This model offers a more intuitive representation of relationship between colours. Basically a colour model is the specification of coordinate system and a subspace within that, where each colour is represented in single point. Hue represents the dominant wavelength in light. It is the term for the pure spectrum colours. Hue is expressed from 0º to 360º. It represents hues of red (starts at 0º),yellow (starts at 60º), green (starts at 120º), cyan (starts at 180º), blue (starts at 240º) and magenta (starts at 300º).Eventually all hues can be mixed from three basic hues known as primaries. Saturation represents the dominance of hue in colour. It can also be thought as the intensity of the colour. It is defined as the degree of purity of colour. A highly saturated colour is vivid, whereas a low saturated colour is muted. When there is no saturation in the image, then the image is said to be a grey image. Value describes the brightness or intensity of the colour. It can also be defined as a relative lightness or darkness of colour [11].The HSV values of a pixel can be transformed from its RGB representation according to the following formula:

$$H = \cos^{-1}\frac{1}{2}\frac{(R-G)+(R-B)]}{\sqrt{[(R-G)^2+(R-B)(G-B)]}}$$

$$S = 1 - \frac{3[min(R,G,B)]}{R+G+B} \qquad V = \frac{R+G+B}{3}$$

Once the colour space is specified, colour feature is extracted from images or regions. A number of important colour features have been proposed in the literatures, including color moments (CM),color histogram, color correlogram etc. The Color moment can be used as remedies of user's queries which are semantic in nature. Color histogram is a popular color feature that has been widely used in many image retrieval systems. Color histogram is robust with respect to viewpoint axis and size, occlusion, slow change in angle of vision and rotation. The color correlogram was proposed to characterize not only the color distributions of pixels, but also the spatial correlation of pairs of colors. Compared to the color histogram the color correlogram provides the best retrieval results, but is also the most computational expensive due to its high dimensionality.

3.1.1.Color moments

To differentiate objects based on color, Color moments have been successfully used in many retrieval systems, especially when the image contains just the object. The basis of color moments is that the distribution of color in an image can be considered as a probability distribution which can be characterized by various moments. ie. If the color in an image follows a certain probability distribution, the image can be identified by that distribution using moments. The first order (mean), the second order (variance) and the third order (skewness) color moments have been proved to be efficient and effective in representing color distributions of images[4].

$$\mu_i = \frac{1}{N}\sum_{j=1}^{n} P_{ij}$$

$$\sigma_i = (\frac{1}{N}\sum_{j=1}^{n}((P_{ij}-\mu_i)^2)^{\frac{1}{2}}$$

$$S_i = (\frac{1}{N}\sum_{j=1}^{n}((P_{ij} - \mu_i)^3)^{\frac{1}{3}}$$

Where P_{ij} is the value of the i- the color channel of image pixel j and N is the number of pixels in the image.

A color can be defined by 3 or more values. Here we can use any of the color coding schemes, say HSV. A moment can be calculated for each of these channels. Thus we get nine numbers—three moments for each color channel as color features for each of the image. Thus color moments are a very compact representation compared to other color features. Due to this compactness, it may also lower the discrimination power.

Similarity between two image distributions is defined as the sum of weighted differences between the moments of two distributions.

ie.

$$d_{mom}(H, I) = \sum_{i=1}^{r} w_{i1}\left|E_i^1 - E_i^2\right| + w_{i2}\left|\sigma_i^1 - \sigma_i^2\right| + w_{i3}\left|S_i^1 - S_i^2\right|$$

where (H,I) are the two image distribution components, i is the current channel index (1=H,2=S,3=V), r is the number of channels ,here 3,E_i^1,$E_i^2$are the first order moments of two image distributions, σ_i^1,σ_i^2 are the second order moments of two image distributions, S_i^1,S_i^2 are the third order moments of the two image distributions and w_i are the weights for each moment. Pairs of images are ranked based on d_{mom} values. The images with lower d_{mom} values are ranked high and are more similar compared to those with higher d_{mom} values.

The methodology used to calculate moments is as follows. We first scale all images to the same size for efficiency. Color moments are based on probability distributions. So image size should not change the result of comparison. We calculate the three color moments using the formula defined above for the Query Image. We then repeat the calculations for our database images. Calculate d_{mom}value after giving appropriate weights and rank the images in the increasing order of this value. The images with the lowest d_{mom} values are selected as the result images. In this way, we can use color moments as a technique to compare images based on color. Color moments can be used as the first pass to narrow down the search space before other sophisticated color features are used for retrieval.

3.1.2. Color Histogram

Color Histogram represents the distribution of intensity of the colorin the image. Color histograms are a set of bins where each bin denotes the probability of pixels in the image being of a particular color. It serves as an effective representation of the color content of an image if the color pattern is unique compared with the rest of the data set. In addition, it is robust to translation and rotation about the view axis and changes only slowly with the scale, occlusion and viewing angle [3].

A color histogram H for a given image is defined as a vector
H = {H[1], H[2], . . . H[i], . . . , H[N]}

where i represent a color in the color histogram, H[i] is the number of pixels in color i in that image, and N is the number of bins in thecolor histogram, i.e., the number of colors in the adopted color model.

In order to compare images of different sizes, color histograms should be normalized. The normalized color histogram H ′ is defined as

$$H' = \{H'[0], H'[1], \dots H'[i], .. H'[N]\}$$

where $H'[i] = \frac{H[i]}{XY}$,XY is the total number of pixels in an image.

An ideal color space quantization presumes that distinct colors should not be located in the same sub-cube and similar colors should be assigned to the same sub-cube. A color histogram with few colors will decrease the possibility that similar colors are assigned to different bins, but it increases the possibility that distinct colors are assigned to the same bins, and that the information content of the images will decrease by a greater degree. color histograms with a large number of bins will contain more information about the content of images, thus decreasing the possibility of distinct colors will be assigned to the same bins.

Minkowski-form distance metrics [12] compare only the same bins between color histograms and are defined as:

$$d(Q, I) = \sum_{i=1}^{N} |H_Q[i] - H_I[i]|^r$$

Where Q and I are two images, N is the number of bins in the color histogram (for each image we reduce the colors to N, in the RGB color space, so each color histogram has N bins), $H_Q[i]$ is the value of bin i in color histogramH_Q , which represents the image Q, and $H_I[i]$ is the value of bin i in color histogram H_I which represents the image I.

When r=1, the Minkowski-form distance metric becomes L_1 . When r=2, the Minkoski-form distance metric becomes the Euclidean distance. This Euclidean distance can be treated as the spatial distance in a multi-dimensional space. In this paper, we will use the square root of Euclidean distance to calculate the distance between two color histograms, which is defined as:

$$d(Q, I) = \sqrt{\sum_{i=1}^{N} |H_Q[i] - H_I[i]|^2}$$

The image retrieval using histogram consists of the following stages. First of all Query image is given from the user. Then the histogram of the color image is calculated. Each image added to the database is analysed and a colour histogram is computed which shows the proportion of pixels of each colour within the image. Then this colour histogram for each image is stored in the database. Finally Euclidean Distance from query image to database images is calculated and sorted the distance in ascending order and the top images are displayed on the screen. Thus we can use color histograms to retrieve matching images from the database. It performs well compared to other descriptors when images have mostly uniform color distribution but it has the disadvantages of lack of spatial information and therefore tends to give poor results. If two images have exactly the same color proportion but the colors are scattered differently, then we can't retrieve correct images using color histogram.

3.1.3.Color correlogram

A color correlogram is a table indexed by color pairs, where the k-th entry for (i, j) specifies the probability off inding a pixel of color j at a distance k from a pixel of color i in the image[5]. Let I represent the entire set of image pixels and $I_{c(i)}$represent the set of pixels whose colors are c(i).Then, the color correlogram is defined as:

$$\gamma_{(i,j)}(k) = Pr_{p1\in c(I),p2\in I}\ [p2 \in I_{c(j)}|p1 - p2| = k]$$

Where i, j ∈ {1, 2, …, N}, k∈ {1, 2, …, d}, and | p1 – p2 | is the distance between pixels p1 and p2.

If we consider all the possible combinations of color pairs the size of the color correlogram will be very large (O(N2d)), therefore a simplified version of the feature called the color auto correlogram is often used instead. The color auto correlogram only captures the spatial correlation between identical colors and thus reduces the dimension to O(Nd) [5].

L1 and L2 distance metrics in Minkowski-form distance metrics [12] are used to compare color features of two images. For correlograms, L1 is used in most cases because it is simple and robust. The distance between two images I and I' is calculated as follows:

$$|I - I'|_{h,L1} = \sum_{i\in[m]} |h_{c_i}(I) - h_{c_i}(I')|$$

$$|I - I'|_{\gamma,L1} = \sum_{i,j\in[m],k\in[d]} |\gamma_{c_i,c_j{}^{(K)}}, (I) - \gamma_{c_i,c_j{}^{(K)}}, (I')|$$

The image retrieval problem in color correlogram is as follows. A Query image is given from the user. Then the correlogram of the color image is calculated. Color correlograms of the database images are also calculated. Then the distance from query image to database images is calculated using L1 metric and sorted the distance in ascending order and the top images are displayed on the screen. Thus we can use color correlograms to retrieve matching images from the database.

3.2 Texture

Texture is another property of image which is used in pattern recognition and computer vision. Texture [13] is defined as structure of surfaces formed by repeating a particular element or several elements in different relative spatial positions. The repetition involves local variations of scale, orientation, or other geometric and optical features of the elements. The ability to match on texture similarity can often be useful in distinguishing between areas of images with similar color (such as sky and sea, or leaves and grass)[6].Thus texture analysis plays an important role in comparison of images supplementing the color feature. Texture representation methods can be classified into Structural and Statistical categories. Structural methods are applied to textures that are very regular. Statistical methods, includes characterizing texture by the statistical distribution of the image intensity.

Many Statistical techniques has been used for measuring texture similarity in which the best established rely on comparing values of second order statistics calculated from query and stored images [11]. These techniques calculate the relative brightness of selected pairs of pixels from each image. From these it is possible to calculate measures of image texture such as the degree of contrast, coarseness, directionality and regularity, or periodicity, directionality and randomness. Alternative methods of texture analysis for retrieval include the use of Gabor filters and Wavelets.

Texture queries can be formulated in a similar manner to colour queries, by selecting examples of desired textures from a palette, or by supplying an example query image.

3.2.1. Gabor filter

The Gabor filter is a statistical method that has been widely used to extract texture features [14]. This is the most frequently used method in image retrieval by texture. There have been many approaches proposed to characterize textures of images based on Gabor filters. In most of the CBIR systems based in Gabor wavelet, the mean and standard deviation of the distribution of the wavelet transform coefficients are used to construct the feature vector [15].

The basic idea of using Gabor filters to extract texture features is as follows.

A two dimensional Gabor function $g(x, y)$ is defined as:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\ exp\left[-\left(\frac{1}{2}\right)\left(\left(\frac{x^2}{\sigma_x^2}\right)+\left(\frac{y^2}{\sigma_y^2}\right)+2\pi j w_x\right)\right]$$

Where $\sigma_x\ and\ \sigma_y$ are the standard deviations of the Gaussian envelopes along the x and y direction.

Given and image I(x,y) it's Gabor transform is defined as

$$w_{mn}(x,y) = \int I(x,y) g^*_{mn}\,(x-x_1, y-y_1)dx_1dy_1$$

Where * indicates the complex conjugate. Then the mean $\mu_{mn}$and the standard deviation $\sigma_{mn}$of the magnitude of $w_{mn}(x,y)$

i.e. f=[$\mu_{00,}\sigma_{00,\ldots.}\mu_{mn}, \sigma_{mn}, \mu_{s-1k-1}, \sigma_{s-1k-1}$] can be used to represent the feature of a homogeneous texture region.

The texture similarity measurement of a query image Q and a target image T in the database is defined by

$d(Q,T) = \sum_m \sum_n d_{mn}\,(Q,T)$

Where $d_{mn} = \frac{\left|(\mu^Q_{mn}-\mu^T_{mn})\right|}{\left|\mu^Q_{mn}\right|+|\mu^T_{mn}|} + \frac{\left|(\sigma^Q_{mn}-\sigma^T_{mn})\right|}{\left|\sigma^Q_{mn}\right|+|\sigma^T_{mn}|}$

If f_g^Q=[$\mu_{00,}\sigma_{00,}\mu_{01}, \sigma_{01}\ldots\ldots, \mu_{35}, \sigma_{35}$] denote texture feature vector of query image andf_g^T=[$\mu_{00,}\sigma_{00,}\mu_{01}, \sigma_{01}\ldots\ldots, \mu_{35}, \sigma_{35}$] denote texture feature vector of database image, then distance between them is given by:

$$d_2 = \sum_{i=1}^{d} \frac{|f_g^Q - f_g^T|}{|f_g^Q| - |f_g^T|}$$

The Canberra distance measure is used for similarity expression. In the case of low level texture feature, we apply Gabor filters on the query image and we obtain an array of magnitudes. The

mean μ_{mn} and standard deviation σ_{mn} of the magnitudes are used to create a texture feature vector f_g. Similarly the Gabor filters of database images are also calculated and Canberra distance measure is used for computing the distance between query and database images and the results of a query are displayed in decreasing similarity order. In this way Gabor filter can be used to match images from the database using texture property of the image.

3.2.1 Haar wavelet Transforms

Wavelet transforms provide a multi-resolution approach to texture analysis and classification. The wavelet transform represents a function as a superposition of a family of basic functions called wavelets. The wavelet transform computation of a two-dimensional image is also a multi-resolution approach, which applies recursive filtering and sub-sampling. At each level, the image is decomposed into four frequency sub-bands, LL, LH, HL, and HH where L denotes low frequency and H denotes high frequency.

If a data set $X_o, X_{1,\ldots} X_{n-1}$ …. Contains N elements [9], there will be N/2 averages and N/2 wavelet coefficient values. The averages are stored in the first half of the N element array, and the coefficients are stored in the second half of the N element array. The averages become the input for the next step in the wavelet calculation. The Haar equations to calculate an average and a wavelet coefficient from an odd and even element in the data set are

$$a_i = \frac{(X_i + X_{i+1})}{2}$$
$$c_i = \frac{(X_i - X_{i+1})}{2}$$

For a 1D Haar transform of an array of N elements, find the average of each pair of elements, find the difference between each pair of elements and divide it by 2, fill the first half of the array with averages, fill the second half of the array with coefficients and Repeat the process on an average part of the array until a single average and a single coefficient are calculated. For a 2D Haar transform, Compute 1D Haar wavelet decomposition of each row of the original pixel values and then compute1D Haar wavelet decomposition of each column of the row-transformed pixels. Red, green and blue values are extracted from the images. Then we apply the 2D Haar transform to each color matrix.

We apply Haar wavelet decomposition of an image in the RGB color space. We continue decomposition up to level 4, and with F-norm theory we decrease the dimensions of image features and perform highly efficient image matching. If A is a square matrix and A_i is its i^{th} order sub-matrix where

$$A = \begin{bmatrix} a_{11}, \ldots a_{1n} \\ a_{n1}, \ldots a_{nn} \end{bmatrix} A_i = \begin{bmatrix} a_{11}, \ldots a_{1i} \\ a_{i1}, \ldots a_{ii} \end{bmatrix} \quad i = 1 \sim n$$

F-norm of A_i is

$||A_i||_F - ||A_{i-1}||_F$ and $||A_i||_F$=0, we can define feature vector of A as

$\{V_{AF} = \{\Delta A_1, \Delta A_2, \ldots . \Delta A_n\}$

The similarity between two images is computed by calculating the distance between feature representation of the query image and feature representation of the image in the dataset. We use Canberra distance for distance calculation of the feature vectors.

$$d(q,d) = \sum_{i=1}^{n} \frac{|q_i - d_i|}{|q_i| - |d_i|}, \text{ where}$$

$q = (q_1, q_{2,\ldots} q_n)$ is the feature vector of the query image,
$d = (d_1, d_{2,\ldots} d_n)$ is the feature vector of the image in the database,
n = number of elements of the feature vector.

A feature vector is extracted from each image in the database and the set of all feature vectors is organized as a database index. When similar images are searched with a query image, a feature vector is extracted from the query image and is matched against the feature vectors in the index. If the distance between feature representation of the query image and feature representation of the database image is small, then it is considered similar. Thus we can use Haar wavelet for matching images from the database.

4. COMBINING THE FEATURES

The image retrieval using only single feature such as color moment or color histogram may be inefficient. It may either retrieve images not similar to query image or may fail to retrieve images similar to query image. To produce efficient results, we use combination of color and texture features. The similarity between query and target image is measured from two types of characteristic features which includes color and texture features. Two types of characteristics of images represent different aspects of property. While calculating similarity measure, appropriate weights are considered to combine the features [9]. The distance between the query image and the image in the database is calculated as follows:

$$d = w1*d1 + w2*d2.$$

Here, w1 is the weight of the color features, w2 is the weight of the texture features and d1 and d2 are the distances calculated using color features and texture features respectively. The distance d is calculated for each query image with all images in the database. The image that has a lower distance value is considered the similar image and the results are ranked in the ascending order of d. From the studies, [16] It is seen that the value of the average precisions based on single features i.e. only Gabor texture features or only Color moments are less than the average precisions of combined features of color moments and Gabor texture features. This shows that there is considerable increase in retrieval efficiency when both color and texture features are combined for CBIR. Also it is found that [8] the texture and color features are extracted through wavelet transformation and color histogram and the combination of these features is a faster retrieval method which is robust to scaling and translation of objects in an image.

4. PERFORMANCE EVALUATION AND INDEXING SCHEMES

The performance of retrieval of the system can be measured in terms of its recall and precision. Recall measures the ability of the system to retrieve all the models that are relevant, while precision measures the ability of the system to retrieve only the models that are relevant [8].

$$Precision = \frac{Number\ of\ relevant\ images\ retrieved}{Total\ Number\ of\ images\ retrieved}$$

$$Recall = \frac{Number\ of\ relevant\ images\ retrieved}{Total\ no\ of\ relevant\ images}$$

where A represent the number of relevant images that are retrieved, B, the number of irrelevant items and the C, number of relevant items those were not retrieved. The number of relevant items retrieved is the number of the returned images that are similar to the query image in this case. The total number of items retrieved is the number of images that are returned by the search engine. In precision and recall, crossover is the point on the graph where the both precision and recall curves meet. The higher the number of crossover points better will be the performance of the system.

The average precision for the images that belongs to the qth category (A_q) has been computed by

$$P' = \sum_{k \in A_q} \frac{P(i_k)}{|A_q|} \text{Where q=1, 2……10.}$$

Finally, the average precision is given by:

$$P' = \sum_{q=1}^{10} (P'_q / 1\,0)$$

Another important issue in content-based image retrieval is effective indexing and fast searching of images based on visual features. The feature vectors of images tend to have high dimensionality and are not well suited to traditional indexing structures. Sodimension reduction is usually used before setting up an efficient indexing scheme. One of the techniques commonly used for dimension reduction is principal component analysis (PCA).It is a general and very recognizable method [17] and an optimal technique that linearly maps the input data to a coordinate space such that the axes are aligned to reflect the maximum variations in the data. The QBIC system uses PCA to reduce a 20-dimensional shape feature vector to two or three dimensions [18].

After dimension reduction, the multi-dimensional data are indexed. A number of approaches have been proposed for this purpose, including R-tree [19], linear quad-trees [20]. Most of these multi-dimensional indexing methods have reasonable performance for a small number of dimensions (up to 20), but explore exponentially with the increasing of the dimension ality and eventually reduce to sequential searching. Furthermore, these indexing schemes assume that the underlying feature comparison is based on the Euclidean distance, which is not necessarily true for many image retrieval applications. One attempt to solve the indexing problems is to use hierarchical indexing scheme based on the Self-Organization Map (SOM) proposed in [21].

4. CONCLUSION

This paper investigated various feature extraction algorithms in CBIR. A study of different color and texture features for image retrieval in CBIR is performed. Numerous methods are available for feature extraction in CBIR. They are identified and studied to understand the image retrieval process in the CBIR systems. Studies made on experiment results show that the method based on hybrid combination of color and texture features has higher retrieval accuracy than the other methods based on single feature extraction. Color moments, color histograms, color correlogram and gabor texture are considered for retrieval. It is difficult to claim that one feature is superior to others. The performance depends on the color distribution of images. The combination of color descriptors produces better retrieval rate compared to individual color descriptors. Color moments and color histogram features can be combined to get better results. Color histograms and correlograms can be combined retaining advantages of histograms with spatial layout. Similarly, Texture feature can be combined with color moments or color histogram to get accurate results for image retrieval. From the studies, it is found that only one color feature or texture feature is

not sufficient to describe an image. There is considerable increase in retrieval efficiency when both color and texture features are combined.

REFERENCES

[1] Kenneth R. Castleman, (1996)"Digital Image Processing" . Prentice Hall .

[2] A. Blaser, (1979) "Database Techniques for Pictorial Applications", Lecture Notes in Computer Science,Vol.81, Springer Verlag GmbH.

[3] Ryszard S. Chora´s(2007) "Image Feature Extraction Techniques and their Applications for CBIR and Biometrics Systems" International journal of biology and biomedical engineering Issue 1, Vol. 1.

[4] M. Stricker, and M. Orengo, (1995) "Similarity of color images," SPIE Storage and Retrieval for Image and Video Databases III, vol. 2185, pp.381-392.

[5] J. Huang, et al., (1997) "Image indexing using color correlogram," IEEE Int. Conf. on Computer Visionand Pattern Recognition, pp. 762-768, Puerto Rico.

[6] Deepak S. Shete1, Dr. M.S. Chavan (2012) "Content Based Image Retrieval: Review" International Journal of Emerging Technology and Advanced Engineering ISSN, Volume 2, pp2250-2459.

[7] Fazal Malik , Baharum Baharudin(2013) "Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain", Journal of King Saud University –Computer and Information Sciences Vol 25 ,pp.207 -218.

[8] Manimala Singha and K.Hema chandran(2012) "Content based image retrieval using color and texture "Signal & Image Processing : An International Journal (SIPIJ) Vol.3, No.1, pp.39-57.

[9] Md. Iqbal Hasan Sarker and Md. Shahed Iqbal (2013) "Content-based Image Retrieval Using Haar Wavelet Transform and Color Moment" Smart Computing Review, vol. 3, no. 3, pp.155-165.

[10] MS. R. Janani, Sebhakumar.P (2014) "An Improved CBIR Method Using Color and Texture Properties with Relevance Feedback International Journal of Innovative Research in Computer and Communication Engineering Vol.2, Special Issue 1.

[11] K. Arthi, Mr. J. Vijayaraghavan (2013) "Content Based Image Retrieval Algorithm Using Colour Models" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 3.

[12] Shengjiu Wang (2001) "A Robust CBIR Approach Using Local Color Histograms" Technical Report TR 01-13.

[13] J. Zhang, G. Li, S. He, "Texture-Based Image Retrieval by Edge Detection Matching GLCM", The 10th IEEE International Conference on High Performance Computing and Communications.

[14] A. K. Jain, and F. Farroknia, (1991) "Unsupervised texture segmentation using Gabor filters," Pattern Recognition, Vo.24, No.12, pp. 1167-1186.

[15] YogitaMistry, Dr.D.T. Ingole (2013) " Survey on Content Based Image Retrieval Systems" International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 8.

[16] S. Mangijao Singh , K. Hemachandran (2012) "Content-Based Image Retrieval using Color Moment and Gabor Texture Feature" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 1, pp.299-309.

[17] Julie M. David, Kannan Balakrishnan, (2014), "Learning Disability Prediction Tool using ANN and ANFIS" , Int. J.of Soft Computing, Springer Verlag Berlin Heidelberg, ISSN 1432-7643 (online), ISSN 1433-7479 (print), DOI: 10.1007/s00500-013-1129-0, 18 (6), pp 1093-1112.

[18] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B.Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, (1995) "Query by image and video content: The QBIC system." IEEE Computer, Vol.28, No.9, pp. 23-32.

[19] N. Beckmann, et al, (1990) "The R-tree: An efficient robust access method for points and rectangles," ACM SIGMOD Int. Conf. on Management of Data, Atlantic City.

[20] J. Vendrig, M. Worring, and A. W. M. Smeulders, (1999) "Filter image browsing: exploiting interaction in retrieval," Proc. Viusl'99: Information and Information System.

[21] H. J. Zhang, and D. Zhong, (1995) "A Scheme for visual feature-based image indexing," Proc. of SPIE conf. on Storage and Retrieval for Image and Video Databases III, pp. 36-46, San Jose.

AUTHORS

Shereena V.B.received her MCA degree from Bharathidasan University, Trichy, India in 2000. During 2000-2004, she was with Mahatma Gandhi University, Kottayam, India as Lecturer in the Department of Computer Applications. Currently she is working as Asst. Professor in the Department of Computer Applications with MES College, Aluva, Cochin, India. Her research interests include Data Mining and Image Processing.

Dr. Julie M. Davidcompleted her Masters Degree in Computer Applications and Masters of Philosophy in Computer Science in the years 2000 and 2009 in Bharathiyar University, Coimbatore, India and in Vinayaka Missions University, Salem, India respectively. She has also completed her Doctorate in the research area of Artificial Intelligence from Cochin University of Science and Technology, Cochin, India in 2013. During 2000-2007, she was with Mahatma Gandhi University, Kottayam, India, as Lecturer in the Department of Computer Applications. Currently she is working as an Assistant Professor in the Department of Computer Applications with MES College, Aluva, Cochin, India. She has published several papers in International Journals and International and National Conference Proceedings. Her research interests include Artificial Intelligence, Data Mining, and Machine Learning. She is a life member of International Association of Engineers, IAENG Societies of Artificial Intelligence & Data Mining, Computer Society of India, etc. and a Reviewer of Elsevier International Journal of Knowledge Based Systems. Also, she is an Editorial Board Member of various other International Journals. She has coordinated various International and National Conferences.

INTENTIONAL BLANK

STUDY OF DEFECTS, TESTCASES AND TESTING CHALLENGES IN WEBSITE PROJECTS USING MANUAL AND AUTOMATED TECHNIQUES

Bharti Bhattad[1] and Dr. Abhay Kothari[2]

[1] M.tech in Software Engineering under guidance of Abhay Kothari
bhartibhattad118@gmail.com
[2]Department of Computer Engineering

ABSTRACT

Testing is the one of the important component of any software engineering process. As we talking about the software's applications then web application is the fastest growing application now a day. So web application or web sites will be tested accurately and correctly. Web testing includes testing of various applications like configuration control, navigation control, state, database etc. Web site testing ensures that there will be no broken links, no images will be missed, there should be no spelling mistakes, no any errors or bugs in software, and the download time should not be so delay as specified. Timeliness, structural quality, content, accuracy and consistency, response time and latency and performance are the major web site's quality factors. Functional, browser, performance, security, usability, database etc testing are performed on any website to make it defect free. Also for any project we also need to maintain the database. So database plays very important role for every organization, so for better results testing of database is required. It is now not only the necessity of project or web application itself but of the organization also to avoid any future problems that can be come in application. As a minute fault in data base can causes data loss that may be uncover able in future. Many tools and frameworks are available for testing of databases or generate test cases to check the applications. When we test the website or any web application and there is difference between expected results and the actual results, there is defect. Defects can be classified in to 3 categories: Wrong, Missing and Extra. Errors can be classified according to priority or severity. According to the severity and priority of the defects, these can be fixed before deliver product to the client. In this paper we represent that on which we can apply tests in on database .how we can perform testing on database. We have also computed the coverage of design of test cases to maintain the quality of testing. By this, we can decrease the time, memory and cost of project to some extent, there by easing the tester to manage their testing phases easily.

KEYWORDS

Manual/Automation Testing, Defects, Defect types, Test cases design, Database testing

1. INTRODUCTION

Testing is very important for any software which is developed by any organization. This is one of the major activities of any software development life cycle as software consists of number of program to perform some different or specific tasks. Software Testing ensures that software which will be deployed to any client it will be error free .Testing is used to demonstrate the

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 79–89, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4907

validity of software at each and every stage. As software testing directly affects the quality of software. Generally large organizations spend 60% of their development cost on testing [11]. Software testing is the phase in which testers identify defects, defect is any variance between actual and expected output. Software testing is the phase for analyzing the software for the purpose of identifies the errors so that the software become error free and clients not get any defects at their own side. This directly deals with the quality of software. Database is one of the important entities of any project also database applications are very complex to handle as it consists of Many components and layered architecture. Testing of database is very essential that database should accept correct data or it perform functions correctly and it must perform acceptable performance when deployed .In this paper we include necessity of software with its types, web site quality, test cases design, and types of testing for web site, defect and its life cycle, database testing, and at last it concluded with when to stop testing and also which defect has been removed on the basis of priority and severity of defects. In this we include the importance of database testing. There may be number of approaches are there to test the database application but currently we discussing about two approaches. With the first approach, application developers carry out their tests for databases on their local database. But this approach cannot fulfill the needs of all the testing stages, especially performance and scalability, due to the limitation of small size of data and test cases. Further more data in local development databases may not be accurate or close to real data as all data will create by developers so they know all valid and invalid conditions but data will be tested on user's point of view. Approach, applications are tested over live databases. But this approach can't be used always as security point of view like high risk of secure data disclosing [11]. Testing of database is different from testing of structural programs as access of database takes through SQL statements so it's important to do right testing for database and rest of codes. A database involves inputs like user's side input is given and instances of databases. In addition to check the expected output with original output some following things has to be checked like:

1. Consistency of database.

2. Reflect database to original environments [11].

Normally in database testing ,as in theory a test run does not fail, if all request and responses are right and database states correct after execution of test run.

1.1 Necessity of software testing:

To find that that software is performing what it supposed to do and software not performing what it is not supposed to do. Basically software testing is doing to find the defects as soon as possible and get them fixed that means it's not all about to correct the code but to search the bugs or find the errors in code in starting phase and try to correct them in shortest possible time. If tester is not got any error then it indicates that testing process required improvement of test cases. So it's the necessity of testing process that the testers must write accurately test cases so that it can find out hidden defects from the programs.

1.2 Types of software:

Software development takes places so many types of applications are build up like desktop applications web applications etc. but web applications are the fastest growing application as we concern about the software application .various applications are coming in market or can be downloaded from play stores so these deploy over internet so that many users are get aware of these and can be use them easily so these must not contain any error. Therefore web application will be tested with maximum accuracy. So web testing includes various tests like load control,

navigation control, configuration etc. web site testing says that there not be any link which is broken , there not be any missing data , values or images n, no spelling mistakes will be there , no bugs will be found etc must be taken in to account.

1.3 Web site Quality:

As there are many sites are launching day by day, and many application are performing or depending on the web sites so there must be maintain the quality of the website so that batter results can be found out. As there are many websites available user s will quickly move to different website if it is very slow or complex.

For example many e-commerce sites are running on different web sites and on that various transactions are performed on that so if there is any error on these sites then it directly affect the transactions operations ,sales or any other major operation. So there must be some factors on the basis of that these web sites can be tested and quality can be maintained. Miller gives some quality factors like:

Timeliness: All the web pages of any web site must be upgraded time to time or daily according to the requirement.

Structural quality*:*

Content, accuracy, consistency, response time, latency and performance are important factors to maintain the structural quality.

1.4 Software development life cycle models:

The choice of specific development life cycle model depends on many things like availability of clear user requirements at the starting, whether the user is from IT background, whether there are major risks associated with the project, whether the product development includes high end research etc.

Development life cycle model	Features
Water fall model	It uses when SRS document is clear. It is simple and effective when requirements are clear. Monitoring of project is simple as defined results or outputs are already available at the end of each stage.
Prototype model	It uses when requirements are not clear. Prototyping is done at developer's cost.eve
Evolutionary development model	Similar to prototype model to some extent.
Spiral model	In this some financial conditions is to be checked again and again.

2. TEST CASES DESIGN

For right testing process one need to take correct input and makes accurate test cases. Test cases are the set of expected inputs and expected outputs developed for particular objectives. According to IEEE STD 610 (1990) defines test case as set of inputs, execution conditions, and expected

outputs developed for a particular objectives, such as to exercise a particular path or to verify compliance with a specific requirement.

Some important features that one should kept in mind:

- Test cases should covers all essential features be balanced. There is balance between all conditions like normal, abnormal and boundary conditions.
- All testing methods will be balanced whatever use like black box, white box, functional or non functional testing.
- Test cases must be accurate, economical repeatable , traceable, self understandable and self standing that means anyone who use the test cases are get the concepts means it should not be so complex.
- Test case must be according to need of description that to be tested that is it must be specific it should not contain any unspecific steps so unnecessary steps must be avoided.
- A good test case is one which give same result each time no maters that who is going to use it. A test case will be appropriate for developers, testers and environment.

Above are the necessity of good test cases but writing good test cases are the another major concept so for writing good test cases some of important things are:

- Improve testability of test cases: testability means easy to test. Test steps will be written in active cases so that all things will be clear and understandable to every person who will going to execute it.
- Improve accurately: accurately means testers follow the test directions, the result of test case pass or fail will be right.While designing the test cases some common mistakes has been done by testers that must be avoided like:
- Don't make too long test cases. Don't combine two or more test cases into one as a single test case may contain many results or verify many criteria.
- Incorrect, in appropriate test cases causes many confusions to testers or the person who going to test.

3. TYPES OF TESTING FOR WEBSITE:

Functional testing: In functional testing various functionalities of the web pages are get tested. It includes various check links, web forms, session testing, css tags, and dynamic contents testing for web pages. Check links includes test a link of a page to other external links, means all links given on the web page will be work properly and accurately. Testing of web forms include for the default values assigned on the web pages, password field for login purpose not show its content that it must be hide from other users, fake input values not be entered or taken for each and every field of the web page. Session testing shows that session cookies must be reset between browser session etc .in this various html ids, attribute all must be identified. It also includes database checking that is data consistency must be maintain, various create ,alter etc commands works properly, data should be give correct data on data retrieval option given by user. Data connectivity must be maintained throughout the session.

Browser compatibility testing: In browser testing various functions check like web application will work on several browsers or not that is it will be compatible to all the browsers equally and correctly, also check browser's security checking like hacking etc.web application must be compatible to operating environment that is check for user interface, desktop integration functionalities.

Performance testing: In this web application can be checked for load and stress that is linear scalability it means web application's performance does not vary as the number of users increases. Load testing identifies scalability index for web application performance. Also check for response of server to browser when submitting request. Also check response of server over different time period. Stress testing specifies how application responding if level of load is very high.also checks the parts of web application if fails under heavy load. Also check for the functions that are responding if load is there and application get fails.

Security testing: As security is another major issue of the web application or any web site. This includes check for different URL's basic authentication should be check, check for invalid inputs and various text fields, check for web site protection of inaccessible web files or directories, check for insecure page for web sites etc should be checked.

Usability testing: This gives user interface with the website; it has large impact on the website. It includes test for navigation control that is page will be easily movable from one page to another, user find instructions for each functions that they require to operate, consistency must be maintain on each page that is from first to last, function should be work properly for each page. Apart from navigation it also includes content checking that is spelling errors must be avoided, content must be arranged logically and correctly, content must be easily understandable for the users, pattern style must be correct that is color, alignment, border, guidelines, frames, fonts etc must be checked.

Production monitoring: This check that web application must be test timely and save the service level agreement. Also check time to time end user experience so that more quality can be maintained. Also check for correct functioning of the application from various geographical locations.

3.1 Database testing:

For any project we need to maintain the data in data base. Database applications play an important role in nearly every organization, yet little has been done on testing of database applications. They are becoming increasingly complex and are subject to constant change. They are often designed to be executed concurrently by many clients. Testing of database application hence is of utmost importance to avoid any future errors encountered in the application, since a single fault in database application can result in unrecoverable data loss. Many tools and frameworks for performing testing of database applications has been proposed to populate the test database and generate test cases which checks the correctness of application. They check database applications for consistency constraints and transactions concurrency. In this paper we present a DBGEN- database (test) GENerator, an automated framework for database application testing. In this framework Test Strategies for testing of embedded SQL queries within imperative language are presented. Finally we present strategies for performing efficient regression tests by reducing the resets that may occur while testing database applications. We have also computed the coverage of design various test cases to predict the quality of testing. By this, we reduce the testing time and cost by approximately by 30%, thereby easing the tester to manage his testing activities easily.

So database plays very important role for every organization, so for better results testing of database is required. It is now not only the necessity of project or web application itself but of the organization also to avoid any future problems that can be come in application. As a minute fault in data base can causes data loss that may be uncover able in future. Many tools and frameworks are available for testing of databases or generate test cases to check the applications. In this paper we represent that on which we can apply tests in on database. We have also computed the coverage of design of test cases to maintain the quality of testing. By this, we can decrease the

time, memory and cost of project to some extent, there by easing the tester to manage their testing phases easily.

Negative and Positive testing: This testing is we give invalid inputs and receive an error that means if database system get any wrong data as input it must give wrong data in response (fig.1). Same way positive test is done that is to give correct input and expect some steps to be completed in accordance with the specification (fig 2).

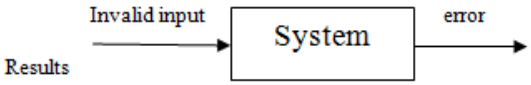


Fig. 1 negative testing

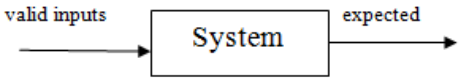


Fig. 2 positive testing

3.2 Types of testing in database:

The figure says the areas of testing involved during different databases testing approaches like black-box testing and white box testing

Black Box testing in database:

In black box testing only testing of output takes places that is not about coding or implementation so in same way black box testing involves interfaces and integration of database, which includes:

- Testing of meta data that is how data is mapped.
- Verifying input data and corresponding output data.
- Checking of data from query languages.
- Various methods used like error guessing, equivalence class partitioning, cause effect graph etc can be used.

For black box testing programmer has good knowledge of designing of database.

White Box Testing in database:

White box testing basically deals with the internal structure of database. Only testing of output takes a place that is not about coding or implementation so specification details are hidden from user.

- Testing of database triggers , views, etc takes places as they are used for database refactoring.
- Verifying different database functions, views, sql queries etc.
- Checking of database tables, data models, schema etc.
- Checking of database consistency and integrity constraints.
- Various methods are used like condition coverage, decision coverage, statement coverage, cyclometric complexities etc.

For white box testing internal knowledge of database is needed due to this internal bugs can be removed from database. But sometimes in this all SQL statements are not covered.

White Box Database Application Technique:

It may be noted that generation of test cases is very important in any testing so while creating test cases for database testing, the semantics of SQL queries must be reflected in all test cases.

So to make this effectively a method is used called as "White Box Database Application Technique".

In this method, SQL queries are covered into GPL statements, followed by traditional white box testing to generate test case that include SQL semantics.[4]

3.3 Database testing cycle:

In database life cycle there are various stages like set fixture, test run, verification of output and tear down.

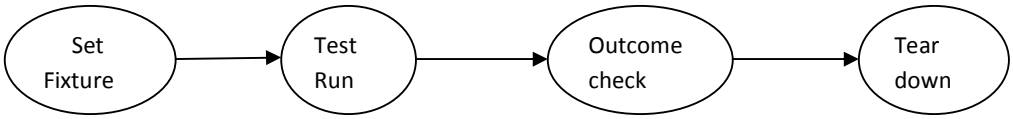


Fig. 3 database testing cycle

In set fixture, is the starting step of database before we start the database testing. Till now test cases have been made so in next step we can test the test cases that are defined for the database. now for this various methods are used n depending on the results we can decide that test cases should be modified or kept as it is and last is tear down stage that results in either terminating testing or start with other test suite(fig.3).

Some problems in database testing:

Major problem of database testing are the cost and size of database as set up for database testing is very high in terms of cost and as size is considered then its complex to maintain the whole database. Sometimes extra overhead is occurs when transaction testing takes places.

3.4 Automated testing:

Tools available for website testing: There are various tools are available for testing the software at different levels like Jmeter, Grinder, Multi-Mechanize, Selenium, Capybara, Capybara, Pylot, Open STA, Web load, Webrat . There are various tools are available for performing database testing.

AETG: it generates test cases from requirement model. It uses combinatorial design methods to find minimal test case sets to cover all input values.

Data Factory: It is data generator tool and works as a data manager for database testing. Its fast and easy source of data .

Data Generator: It is automated data generation tool which enables you to create test data for assurance of software quality and performance of database which includes database load, stress and endurance of database. It usually generates random test data for purpose of various types of testing like system, integration, module etc.

Test Data Generator: It used to generate data, tables (views, procedures etc) for database testing.

Datatect: It is powerful software tool for generating varieties of realistic data to ASCII files and RDBMS includes Oracle, Sybase, and SQL Server.

Jenny: It is tool used for carrying out regression testing properly. Always exhaustive testing or condition of endurance testing is creates problems so it covers most of the test cases.

4. TESTING CRITERIA

There are some criteria should be decided about when to start and stop the testing.

4.1 Testing starting criteria

There is another issue that when to start the testing. So for this, Timing is the major factor that as soon as we get software requirements or baselines we can start testing because incorrect requirements results wrong design and implementation and after implementation has been done it becomes very difficult and also expensive to find the defects and correct them. Every project has some requirements and this has two perspectives; one is from user importance and other is according to user usage, depending o\upon these two characteristics a requirement can be generated and a plan or strategy can be made which also means that estimates change accordingly. So objective for starting testing is to trap the requirements related defects as early as possible.

4.2 Testing completion criteria

As 100% testing is not possible and also it's not possible to make the entire defect free software but we need to consider some parameters or factors on the basis of that we can stop the testing process like

- Deadlines as release deadlines, testing deadlines of the project,
- Test cases completed with certain percentage passed .
- Alpha and Beta testing period ends or as we reach optimum level of testing.

5. DEFECT

Defect is any flaw in the software system. When we test the website or any web application and there is difference between expected results and the actual results, there is defect. Defects can be classified in to 3 categories: Wrong, Missing and extra (table 1). There are many examples of defects that can come in testing the website or any web application like:

Table 1: Categories of defects

Category	Examples
Wrong	▪ User gives wrong/incomplete requirements for developing web application. Error in coding, in testing, Data entry errors also Mistakes in error correction or Analyst interprets requirements incorrectly
Missing	▪ Incorrect design specifications or missed out some any specifications.
Extra	• Developer done something extra in web application but client doesn't require. Poor documentation.

5.1 Defect life cycle:

In defect life cycle there are various defects like new, open, review, rejected, test verified, not a defect etc (fig 4). Defect age or phase age is the important concept in testing that means later we find the defect the more it cost to fix it. Defect spoilage is the concept which works on same concept that how late we find the defects or bugs. When defects are getting fixed during defect life cycle then Retesting and Regression testing is performed. Retesting is testing is performed to check that defect get fixed or not while in regression testing is performed to check that checked tests should not affect the unchecked tests.

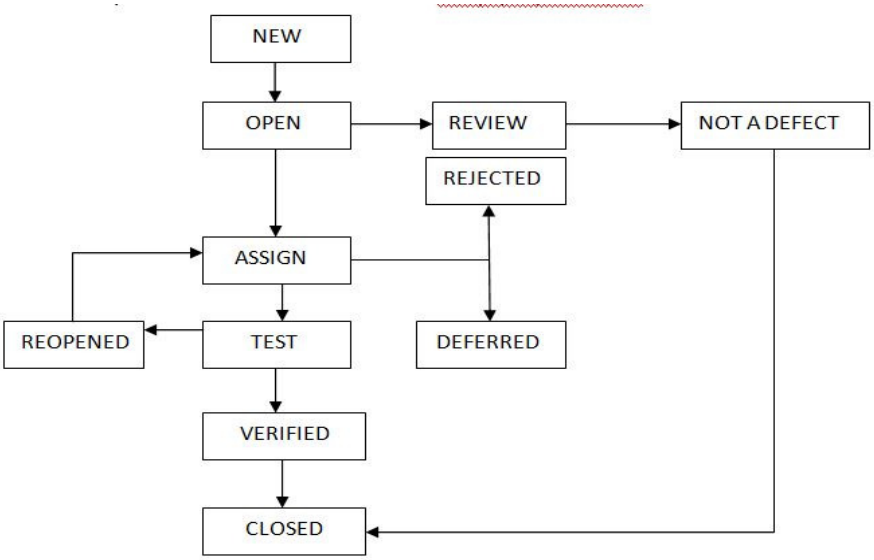


Fig. 4 States of defects

5.2 Severity of defects:

Severity shows how bad the bug is and reflects its impact to the product and to the user. This changes from organization to organization or varies from projects to projects (table 2).

Table 2: Severity of defects

Severity	Description	Criteria
1	Very high	Inability to install/uninstall the product, product will not start, product hangs or operating system freezes, data corruption, product abnormally terminates etc so they are also called showstopper defects.
2	High	Function is not working according to specifications, critical to customer etc that means application can continue with severe defects.
3	Medium	Incorrect error message, incorrect data, etc means application continue with unexpected results.
4	Low	Spelling, grammar mistakes etc that is defects with these severities are suggestions given to client to make application better.

5.3 Priority of defects:

Priority can be decided on the basis of how frequently the defect occurs i.e. probability of occurrence of defect (table 3).

Table 3: Priority of defects

Priority	Description	Criteria
1	Very high	Immediate fix, block further testing
2	High	Must fix before product is released
3	Medium	Should fix if time permits
4	Low	Would like fix but can be released as it is

6. CONCLUSION & FUTURE WORK

Testing is assurance of quality of any kind of software so in order to develop the quality software or web project testing is an essential phase of any development cycle. Testing is widely used now days to help the developers to make the defect free software or any web applications. Although there are so many testing techniques are there but still main important things to do the proper defect management of the defects that occurs. According to the severity and priority of the defects, these can be fixed before deliver product to the client. Defect age and defect spoilage is also important concepts about defect management in web site implementation or web applications as to make the web site error free in optimum level of testing. In this paper, investigation of testing methodologies, and in addition of that focus on test cases design, database testing and defects categories for web based projects are discussed. In near future, the presented literature is utilized to develop a web testing tool.

REFERENCES

[1] "What Is a Good Test Case"Cem Kaner, J.D., Ph.D. Florida Institute of Technology Department of Computer Sciences kaner@kaner.com STAR East, May 2003

[2] Software testing best practices,Ram Chillarege,IBM Research-Technical report RC 21457

[3] Software Testing Techniques- Shivkumar Hasmukhrai Trivedi [B.com, M.Sc I.T (Information Technology), P.G.D.B.M –Pursuing] Senior System Administrator, S.E.C.C [Socio Economic and Cast Census] Central Govt. Project – Bhavnagar [Gujarat – India], Volume 2, Issue 10, October 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[4] Software Testing Research and Software Engineering Education-Thomas J. Ostrand,Elaine J. Weyuker, FoSER 2010, November 7–8, 2010, Santa Fe, New Mexico, USA..Copyright 2010 ACM 978-1-4503-0427-6/10/11 ...$10.00.

[5] Software Engineering by Roger Pressman

[6] Test Case Prioritization Techniques, Siripong Roongruangsuwan, Jirapun Daengdej, Journal of Theoretical and Applied Information Technology© 2005 - 2010 JATIT & LLS. All rights reserved.

[7] Test result reporting by Indira R(white paper of Infosys)

[8] A Study on Software Testing ,H.S. Samra ,Volume 3, Issue 1, January 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[9] A parametric approach for security testing of internet application, Arun K Singh,Anand J Iyer, Venkat Seshadri (white paper of Infosys)Enabling Performance Testing with functional testing tool,by Dick Van Driel,Surya Prakash(white paper of Infosys)

[10] International Journal of Database Theory and Application Vol. 2, No. 3, September 2009 27 DBGEN-Database (Test) GENerator -An Automated Framework for Database Application Testing 1Askarunisa A., 2Prameela P, and 3Dr. Ramraj N 1,2Thiagarajar College of Engineering, Madurai, Tamilnadu, India

[11] http://sqa.fyicenter.com/FAQ/Testing-Tools/Database_Testing_Tools.html

AUTHORS

Ms. Bharti Bhattad

I completed B.E. in computer science from Indore Institute of Science And Technology and pursuing M.tech in software engineering. I also completed diploma in testing from seed InfoTech , Pune. I published some papers on Software Testing. My areas of interests are software testing, Manual testing, test cases design, bug reporting etc.

Dr. Abhay Kothari

He completed Ph.D., MS and B.E..Currently he is working as Professor at Acropolis institute of technology and research. He has 23 years of Working experience. His 18 International and 15 National papers are got Published he get award He get award of IT Excellence by Ministry of Information Technology.

INTENTIONAL BLANK

AD SHARING IN SOCIAL NETWORKS : ROLE OF USER DEFINED POLICIES

Venkata N Inukollu[1], Sailaja Arsi[1] Divya D Keshamoni[2] and Manikanta Inukollu[3]

[1]Department of Computer Science Engineering, Texas Tech University, USA
{narasimha.inukollu, sailaja.arsi}@ttu.edu
[2]Rawls College of Business, Texas Tech University, USA
divya.keshamoni@ttu.edu
[3]Department of Computer Science, Bhaskar Engineering College, India
mani.inukollu@yahoo.com

ABSTRACT

Security policies describe the demeanor of a system through specific rules and are becoming an increasingly popular approach for static and dynamic environment applications.Online social networks have become a de facto portal for Internet access for millions of users. Users share different content on social media sometimes which includes personal information.However, users entrust the social network providers with such personal information.Although social networking sites offer privacy controls, the sites provide insufficient controls to restrict data sharing and let users restrict how their data is handled and viewed by other users.To match the privacy demands of an online social network user, we have suggested a new security policy and have tested the policy successfully on various levels.

KEYWORDS

User defined policy, Social networking, Policy, User privacy.

1. INTRODUCTION

Security policy can be stated as what it means to be secure for a system, or an entity or an organization.The security policy should protect both people and information. Moreover, a security policy helps by minimizingthe risks and by compliance tracking with regulations and legislations.

The policies are used by different employees at different levels of the company. The classification can be done by different levels, such as management level, technical staff level and the end user level. Therefore, people play a vital role in explaining, maintaining, monitoring and using the policies. Thus, giving the user/consumer a high responsibility in defining a policy according to the user requirements would be appropriate for a successful product.

As per the survey results [10], 90% of the people are concerned about issues such as security, privacy, create fake accounts with once identification, children care, and people following them in their social networks. Of all the issues security and privacy is the major concern for nearly

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 91–98, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4908

~70% people. Online social networking sites are the most widely used means of communication. Thus, social media has become an integral part of modern life and so has sharing online content. 59%of people reported that they frequently share online content with others (Allsop, Bassett, and Hoskins 2007), *the New York Times story* receives a tweet once every four seconds. There are several advantages and disadvantages of using social networking. The main disadvantages of using any online social networking sites are:

a) users lose some privacy compared to not being on a social networking site.
b) users may take a step back while sharing some content such as advertisements.
c) Some users want to sustain a line between private and public life, but through social networking, the user may regret later for posting pictures that the user had thought were funny at that minute.

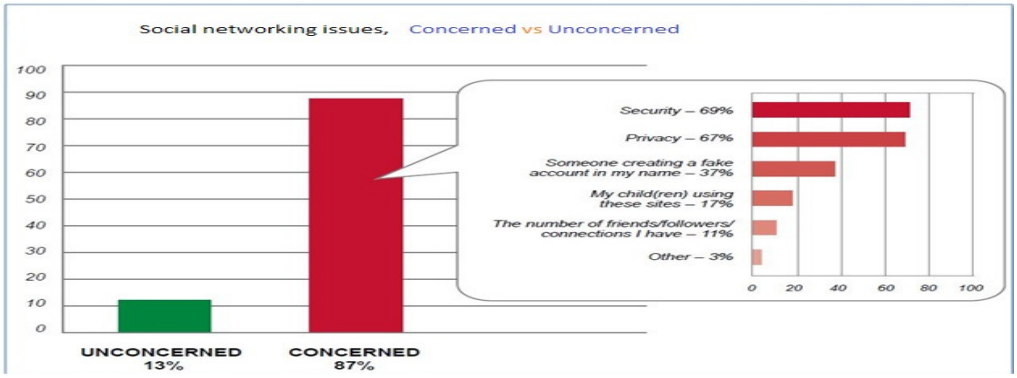


Figure (1): Concernd vs Unconcernedstatistics

So, there should be a common dynamic policy and that should be handled by the user. As the user is given the prime responsibility, user can maintain security and privacy accordingly.

Advantages of user defined security policies:

- user defined security policies address a wide variety of aspects of system security, such as access control, authentication, availability and auditing.
- user defined security policies provide a clear perspective of how to use data abstraction.
- user defined security policies can be applied to distributed and dynamic environment systems.
- user defined security policies providecustomer satisfaction by maintaining privacy and security.
- user defined security policy is a well defined communication paradigm.
- In user defined security policies, creation of object, subject, constraints and actions are all dynamic in nature and follow user satisfaction.
- user defined security policies, can resolve policy conflicts, as user defined security policies maintain separate entities belonging to their domains. User defined security policy has a clear perspective of how entity attributes are defined and exploited.

2. MOTIVATION AND RELATED WORK

An immense amount of work has been done along the user defined privacy by Randy Baden et al. [4], the work is focused on online social networking. In this paper, the authors have introduced a new model called as persona which acts as an online social network that puts policy decisions in

the hands of the users. Persona includes attribute-based-encryption, traditional public key cryptography, and an automated key management mechanism to translate between the two cryptosystems. The persona is a well-defined model to achieve privacy and also provides user-defined access controls needed in online social networking.

Paul Ashley et al. [5], defines a fine grained privacy model called as Platform for the Enterprise Privacy Practices (E-P3P). The E-P3P privacy policy defines, on certain data categories the actions to be performed by the data users. This paper has given a well defined constructed model and semantics for E-P3P. The model comprises of formal semantics of simple requests, compound requests, language details, deployment mapping, defining conditions, authorization algorithm, and related interfaces. This paper has strongly stated that formalized and strictly enforced privacy practices enable enterprises to provide the level of privacy promised using privacy statements.

Oracle 10g Enterprise Manager had released a step by step example for creating user defined policies and metrics [6]. Oracle Enterprise Manager in its own way created and modified tables to provide a simplified user defined policy and user defined metrics that has a user defined policy group interface.

Dasan in [7] explained an automatic method and system for retrieving information based on a user-defined profile such as personalized newspaper, wherein the user defined profile identifies information which is of interest to the user. Dasan has provided well-defined models for user-defined profile systems.

Schreckling et al. [8] introduced a real-time monitoring and enforcement framework for Android called Kynoid. Kynoid is based on user-defined security policies that are defined for data-items. The article on Kynoidmainly focuses on android framework, but the article does not mention whether kynoidcan be applied to different fields, frameworks/ platforms.

3. ROLE OF USER DEFINED SECURITY IN AD MARKETING

Consumers signal their identities through brands and products [11,12], or even restore their original self-view through brands [13].Companies often create online ad campaigns or encourage consumer-generated content in the hope that people will share this content with others, but some of these efforts take off while others fail.Despite the fact that viral marketing and advertising can be a successful means of marketing communication, there is still a limited understanding of how it works and why consumers' share online content[14].

Previous research indicates that consumers tend to share advertising messages that they find entertaining, informative, titillating, or shocking-that is, messages that evoke strong emotional responses [14].This research proposes that among various factors , Privacy controls for sharing and receiving content for different types of products(Public vs Privately consumed products) are important drivers of social transmission of information on the social media.

Publicly consumed products are those that are seen by others when being used, privately consumed products are ones that are not seen during the consumption process with the possible exception of the user. Research by [15] suggests that if consumers can be assured of their privacy, firms can use personalization of ads to generate higher click-through rates. According to [16], knowing that our consumption decisions are going to be subject to public scrutiny will influence our consumption choices. Similarly, individuals might be concerned about the information / advertisements that they share online because of the desire to be evaluated favorably by others. Individuals might sometimes feel compelled to switch away from sharing

advertisements/content (of privately consumed goods). For e.g. hair fall products, foot fungus cream) because of how they expect to be perceived by others. Hence it can be proposed that control over personalized sharing and receiving of particular ads (especially for products linked to private consumption) increases the likelihood to share ads. Hence secure sharing plays a vital role in business development.

Eg: Presently Facebook displays several advertisements on the right hand corner of the page and users do not have the authority to share those advertisements. Also videos shared on Facebook by an individual are available publicly and potentially to all the friends in the users list. The current research examines if personalized sharing and receiving advertisements in private will boost individuals to partake in more number of advertisements online especially on Facebook. For e.g. If an ad about a fairness product appears on my fb page and if I would relish to share this ad only with one of my friends who is withal probing for more preponderant fairness products and if both of us have control over sharing/receiving the information without other friends on our face book knowing about it,it would enhance privacy.This would increase the number of ads clicked and shared.

4. USER DEFINED SECURITY POLICIES

One of the key foundations of a secure computer system is the security policy [1]. A security policy is a set of objectives, rules of behavior for users and administrators, and requirements for system configuration and management that collectively are designed to ensure security. Security policies can be expressed by associating security labels with either the data or actions that protect the system, as defined by Swamy et al. [2]. Based on the available literature, the following observations were made:

a) For each system's environment, assigning an individual policy to the system would be inappropriate because, for a particular system, there can be different environments, such as login authorization module, privileges module, malware defense module, account monitoring and control module, and testing module and for each environment it is not possible to have an individual policy applied to it.

b) For the entire system, the specification languages use only the common policies such as authorization, prohibition, obligation, delegation, information filtering, and refrain policies. All the environments within the system are explained using those common policies. There is no scope for policy extendibility.

From the above observations, user defined security policies are an arena that has not been handled and which needs great attention. User defined security policy means defining and implementing the system's scenario, according to the user. The user is afforded a main/primary role and the user takes the responsibility to ensure the security and privacy of the system. Currently, in that respect is the rapid growth in computer and online network resource utilization.In these circumstances, privacy is of top priority to preserve.

If users have a policy implementation scope, a user can develop the system according to the requirements while ensuring privacy. If such a user defined security policy comes into existence, then there will be many advantages for businesses, industries and organizations.From the table, if observed, there were several instances which were not handled by the existing policies. For those instances/environments we can assign the user defined security policy and let users implement their own policy, which handles both security and privacy of the system.

5. SECURE AD SHARING FRAMEWORK

Ad marketing plays a vital role in social networking application in the context of revenues. The content on the social network consists of various kinds of data, such as personal, professional, and location based, but the current frameworks don't provide much needed privacy and security for the personal data. The only private communication that exists in the present social networking system is through mail Communication. In order to provide more privacy and security, we introduce a new component in the existing framework as user defined view where a user can have more control on his/her personal data. And we also introduce the new component as secure share where user can share his content with specific people or group of people. Currently social network applications provide sharing of information which can be viewed by public.(other than mail communication).

Figure (2) shows the social network application framework with the new components. User network shows all the data associated with the user account, such as photos, videos, social network, and text messages. In this level we introduce a new component called secure share where a user can share his/her private/personal information with specific people or group of people. All the user activities are carried out by the pre-defined access control polices in the social network system. In the view level, we have introduced one more new component called user-defined views, where user can see their own personal data, Shares (such as posts, messages, mail communication, multimedia sharing, ad sharing) that are shared by other users in his/her social network.

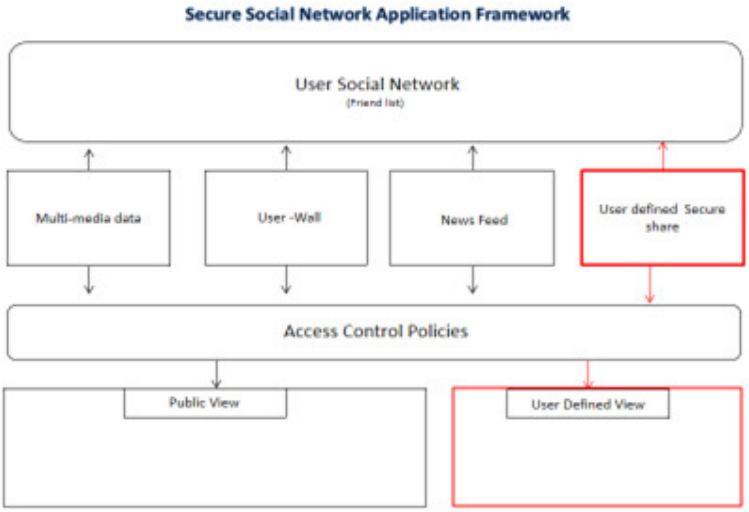


Figure (2): User defined components in Social network

6. CASE STUDY: FACE BOOK

Online social networks have become a de facto portal for Internet access for millions of users. Users mostly share private content such as personal information, photographs, gender preference, marital status, political and religious views, and identity of friends, occupations and phone numbers with their friends [3]. However, users entrust the social network providers with such personal information. Although social networking sites offer privacy controls, the sites provide insufficient controls to restrict data sharing and to let users restrict how their data is managed and viewed by other users. In this aspect, social networking sites lack the privacy of users. On high profile web sites the leakage of personal information [4] has heightened web user's privacy concerns. In that location requires to be an approach to prevent unintended leakage and

manipulation of sensitive data, and that could offer strong guarantees, that deployed applications respect privacy and protection policies. To meet the privacy needs of an online social network, there should be a policy that puts the decision in the hands of the users which is addressed as 'user defined security policy'.

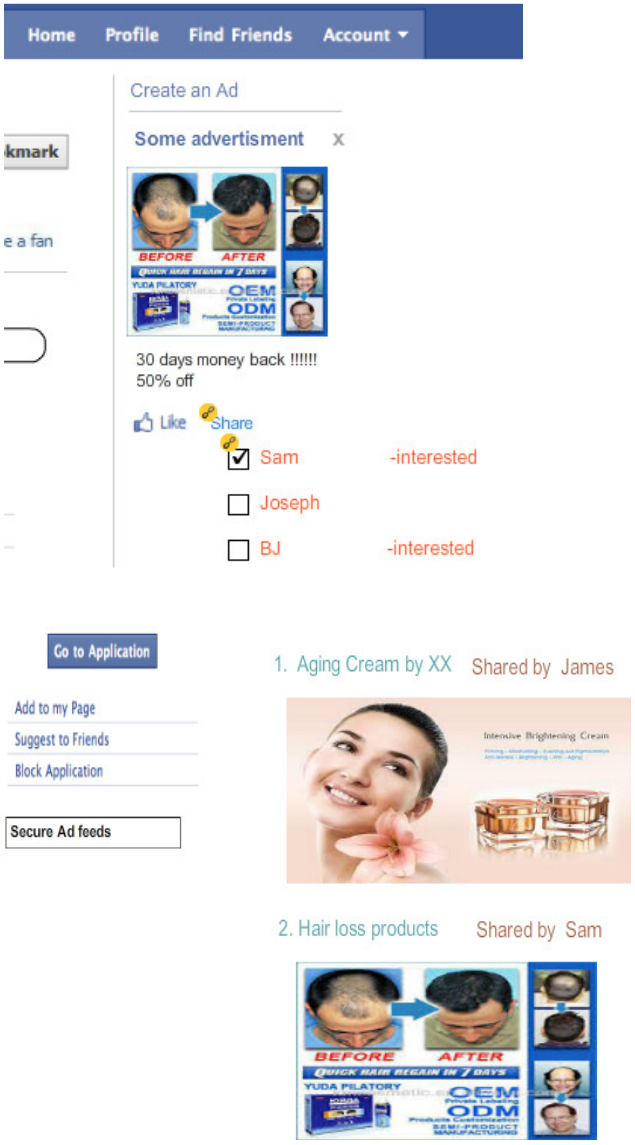


Figure (3): Facebook plug-in

As privacy is becoming a competitive advantage, enterprises need to secure user's personal information against both malicious privacy breaches and unintended data flows due to carelessness [3]. The main goal is to restrict the information shared with friends (in social network terms) to what might be appropriate.

In Facebook when an individual desires to share something, the options are limited, i.e. a user cannot share the data securely and thus this leads to exposure of user's personal interests/information such as the individual's interest in sharing a particular type of content in this instance of communion.In figure(3), we have implemented in face book , where a user has the

freedomto share the data only with desired people and those contents can be viewed in secure ad feeds as shown in the figure(3), and this entire process is carried out in the user defined view.

7. EXPERIMENTAL RESULTS

We have implemented these components in the Facebook social network plugin and conducted a survey to validate our propositions.To our surprise every user in the survey liked the concept of user defined view and securelyshare, which shows that people are more concerned about the privacy and security of their personal and professional data. 70% people voted for ad sharing through secure share and 30% people votedfor personal data sharing through secure share. Additionally, from the survey we observed that the secure sharingof advertisements influences people in terms of buying products.

8. SUMMARY

Social networks have various issues in terms of privacy and security of the users personal and professional data. In order to provide additional security and privacy, we have introduced two new components, namelysecure share and user defined view in the existing social network framework and studied experimental results with the help of face book plug-in. As per the survey, user defined view provides much needed privacy to the users and secure share helps to improve their privacy and social communication. The research thus proposes that Social media users' message-sharing behaviors are motivated by the need for secure sharing.The social process of sharing an online advertising message shapes and helps express the consumers' sense of self, such that it influences which messages consumers are most likely to share with others. Thus, advertisements of privately consumed goods are passed on to friends and peers under shared privacy controls As per the survey results secure sharing exponentially increases the sharing of advertisements on social media and also helps the marketers to reach potential customers.This study contributes greatly to marketing practice as well. The ability to create and replicate successful viral advertising campaigns still remains challenging for online marketers. For every advertisement that successfully generates viral buzz, dozens fizzle. For the most part, Social media marketing practitioners still struggle to exploit an opportunity that has tantalized them for more than a decade. Any new insight into why consumers share some messages, but not others is thus significant and hence this research contributes significantly to both academic research and marketing practices.

REFERENCES

[1] Kuhnhauser W.E., "A Paradigm For User-Defined Security Policies," Proceedings of the Reliable Distributed Systems, 1995, Proceedings of the 14th Symposium, Bad Neuenahr, Sep 13-15, 1995, pp. 135-144.

[2] Swamy N., Corcoran B.J., and Hicks M., "FABLE: A Language for Enforcing User-defined Security Policies," Proceedings of the Security and Privacy, 2008, SP 2008, IEEE Symposium, Oakland, CA, May 18-22, 2008, pp. 369-383.

[3] Preibusch S., "Information Flow Control for Static Enforcement of User-Defined Privacy Policies," Proceedings of the Policies for Distributed Systems and Networks (POLICY), 2011 IEEE International Symposium, Pisa, Jun 6-8, 2011, pp. 133 - 136.

[4] Baden R., Bender A., Spring N., Bhattacharjee B., and Starin D., "Persona: An Online Social Network with User-Defined Privacy," Proceedings of the SIGCOMM'09, Barcelona, Spain, Aug 17-21, 2009.

[5] Ashley P., Hada S., Karjoth G., and Schunter M., "E-P3P Privacy Policies and Privacy Authorization," Proceedings of the WPES'02, Washington, DC, USA, Nov 21, 2002.

[6] Oracle Enterprise manager 10g, "Step By Step Example for Creating User Defined Policies and Metrics".

[7] Dasan V.S., “United States Patent,” Jun 2, 1998.
[8] Schreckling, D., Posegga J., Köstler J., and Schaff M., "Kynoid: Real-Time Enforcement of Fine-Grained, User-Defined, and Data-Centric Security Policies for Android," Proceedings of the 6th IFIP WG 11.2 International Workshop, WISTP 2012, Egham, UK, Jun 20-22, 2012, pp. 208-223.
[9] Aaker, Jennifer L. (1999), "The Malleable Self: The Role of Self-Expression in Persuasion," Journal of Marketing Research, 36 (1), 45-57.
[10] Privacy stastics, http://www.welivesecurity.com/2011/06/22/linkedin-privacy/
[11] Kleine III, Robert E., Susan Schultz Kleine, and Jerome B. Kernan. "Mundane consumption and the self: a social-identity perspective." Journal of Consumer Psychology 2.3 (1993): 209-235.
[12] Laverie, Debra A., Robert E. Kleine III, and Susan Schultz Kleine. "Reexamination and extension of Kleine, Kleine, and Kernan's social identity model of mundane consumption: The mediating role of the appraisal process."Journal of Consumer Research 28.4 (2002): 659-669.
[13] Gao, Leilei, S. Christian Wheeler, and Baba Shiv. "The “shaken self”: Product choices as a means of restoring self-view confidence." Journal of Consumer Research 36.1 (2009): 29-38.
[14] Dobele, Angela, et al. "Why pass on viral messages? Because they connect emotionally." Business Horizons 50.4 (2007): 291-304
[15] Goldfarb, Avi, and Catherine E. Tucker. "Privacy regulation and online advertising." Management Science 57.1 (2011): 57-71.
[16] Ratner, Rebecca K., and Barbara E. Kahn. "The Impact of Private versus Public Consumption on Variety-Seeking Behavior." Journal of Consumer Research29.2 (2002): 246-257.

APPLICATIONS OF DATA MINING IN INTEGRATED CIRCUITS MANUFACTURING

Sidda Reddy Kurakula[1], Lokesh Kulkarni[1], Madhu Dasari[1], Helen Armer[2]

[1]Applied Materials India (P) Ltd, Bangalore, India,
Sidda_Reddy_Kurakula@amat.com, Lokesh_Kulkarni@amat.com,
Madhu_Dasari@amat.com
[2]Applied Materials, Inc., Santa Clara, California, USA
Helen_Armer@amat.com

ABSTRACT

Integrated circuits (a.k.a chips or IC's) are some of the most complex devices manufactured. Making chips is a complex process requiring hundreds of precisely controlled steps such as film deposition, etching and patterning of various materials until the final device structure is realized. Also, each chip goes through a huge number of complicated tests and inspection steps to ensure quality. In IC manufacturing, yield is defined as the percentage of chips in a finished wafer that pass all tests and function properly. Yield improvement translates directly into increased revenues. A humongous amount of data (Terabytes per day) is logged from the equipment in the fab. This paper describes some applications of advanced data mining techniques used by chip makers and equipment suppliers in order to improve yield, match equipment, increase equipment output and also to predict the change in equipment performance before and after maintenance activities.

KEYWORDS

Integrated Circuit (IC), Yield, Equipment, Wafer, Models

1. INTRODUCTION

The process of creating integrated circuits (IC's) is called wafer fabrication. It is a sequence of chemical and photographic steps (like lithography, etching, deposition, oxidation and diffusion) in which the circuits are constructed on a semiconductor material typically called a wafer. In order to perform Automatic Process Control and offline data mining, a large amount of data is collected, stored and retrieved from the equipment in which the said processes are being carried out. In IC manufacturing, yield is defined as the percentage of chips in a finished wafer that pass all tests and function properly. Due to the large number of processing steps and the complex interactions between steps, yield is a complex function to analyse. Data mining methods for yield analysis are now starting to be developed and deployed. Table 1 depicts the yield values for 2 different process technologies with a total number of steps (N) equal to 200 and 450. As can be seen, total perfection at each process step is absolutely necessary for achieving higher yields.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 99–107, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4909

Table 1. Yield for two different process technologies.

x	*Y* for *N* = 450	*Y* for *N* = 200
95%	9.45×10^{-9} %	3.51×10^{-3} %
99%	1.09 %	13.4 %
99.9%	63.7 %	81.9 %

Where

x = Success rate of each process step and
Y= Yield of working devices (i.e. $(x/100)^{N}$ %)

2. DATA MINING METHODOLOGY

In the proposed data mining methodology a holistic approach is being followed by the FabVantage™ group at Applied Materials. This methodology goes beyond traditional statistical process control methods which primarily emphasize process monitoring for change-point detection; instead it focuses more on yield driven control limits and strategies.

2.1. Data sources and typical volumes

Different types of data that are generated and used for various purposes in an IC fabrication unit are event logs data, unit processes data, integration data, inspection & review data, metrology data and parametric and final yield data. The size of each type of data varies from a few gigabytes/day to a few terabytes/day depending on production capacities. "Unit processes" constitute 30 to 40% of total process steps involved in making an IC. All the data mining techniques described in later sections were mostly used to analyse this "Unit Processes" data. The Equipment Data Acquisition is typically performed by one or more factory data gathering or analysis software applications (clients) using different standards like SECS (Semiconductor Equipment Communications Standard), GEM (Generic Equipment Model) or Interface A. One example of this type of client is the Applied Materials E3™. E3 is the only equipment engineering system solution that combines statistical process control (SPC), fault detection and classification (FDC), equipment performance tracking (EPT), advanced data mining (ADM), run-to-run control (R2R) and tool automation on a unified platform. In addition to this most of the Equipment Controller Software provides the option to export/store the data in various file formats. The generic input and output model in data mining for IC fabs is shown in Figure 1.

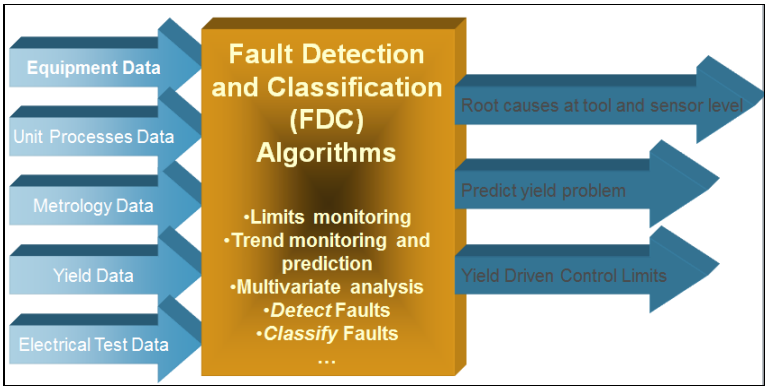


Figure 1. Inputs and outputs of FDC algorithms in IC fabs.

2.2. Data quality check

The unit process data (mentioned above) typically called equipment (tool) sensor data consists of all the physical variables like temperature, flow, pressure, acceleration, torque, angle and the like. This data is recorded during each wafer run. A data quality check algorithm verifies this data for any missing sensors from the defined data collection plan, missing data or stagnant data and also detects any non-optimal sampling rate in any sensor.

2.3. Knowledge Base and sensor priority

Knowledge base is the repository of methodologies, tool documentation, Best Known Methods (BKMs), lessons learned, sensor data collection plans, previously used models and the like. A key part of the knowledge base is sensor prioritization. Each sensor collected from a tool is assigned a priority of P1 through P4 to denote its impact on yield. P1 sensors are known to impact yield if they go out of range, while P4 sensors are known to have no yield impact if they go out of range. P2 and P3 sensors are suspected to impact or not impact yield, respectively, if they go out of range. Sensor priorities are used to reduce the burden of analysing 1000's of sensors in the initial stages of data mining. Table 2 shows an example of sensor priorities.

Table 2. Example Priority Sensor list for one specific tool/process type.

Sensor Name	Sensor Units	Sensor Priority	Sampling Frequency(Hz)
Temperature	degC	P1	1
H_2 Gas Flow	sccm	P1	1
RF Forward Power	Watts	P1	5
Chamber Process Pressure	mtorr	P1	1
Foreline Pressure	mtorr	P2	1
E-Chuck Voltage	V	P2	1
Gas line pressure	psi	P3	0.5
Wafer Counter	None	P4	1 sample per wafer run

Where

P1: Confirmed/known to have caused a yield problem,
P2: Science suggests there will be a problem but no experience from data,
P3: No knowledge whether or not there will be a problem,
P4: Known to be a non-issue.

2.4. Tools and modelling methods

A combination of data mining software including Applied Materials E3, R [1], JMP [2], and UNIX scripting were used to perform the statistical modelling, associated data preparation and reporting. A variety of modelling techniques are used. These include Rules Ensemble [3], Random Forest [4], Support Vector Machines (SVM), Partial Least Squares (PLS) Analysis and Discriminant Analysis in supervised learning and clustering analysis or Principal Component Analysis in unsupervised modelling.

3. ADVANCED DATA MINING APPROACH

3.1. Data visualization

Data visualization is the first and foremost analysis task that helps to identify obvious abnormalities while performing wafer to wafer comparison, lot to lot comparison or recipe to recipe comparison in the tool sensors data. It also helps to rebuild the recipe to quickly compare the recipe under investigation with the BKM recipe. The differences can be summarized and can be used at a later stage while running any classification/regression models as sometimes the differences are deliberately set. Table 3 shows one example recipe wherein the differences from BKM recipe are highlighted in *italics*.

Table 3. Comparison of a recipe with BKM.

Attribute/Step Number	**1**	**2**	**4**	**5**
Step Name	Stabilization	Deposition	Purge	Pump
Step Time (Sec)	2	30	5	10
Temperature (deg C)	200	*250 (BKM is 270)*	190	180
Pressure (Torr)	5	5	2	1
Flow (sccm)	200	*200 (BKM is 250)*	50	10

Customized plotting like shown in Figure 2 was performed in R in order to define the Univariate Analysis (UVA) models using appropriate statistics in different process regions like maximum/slope in transient regions and similarly mean and standard deviation in stable regions.

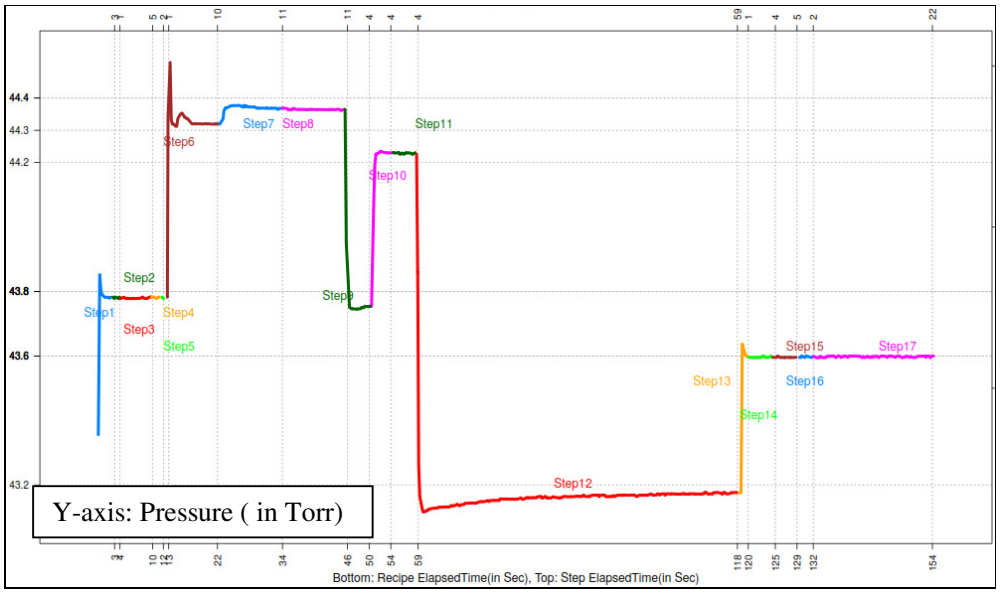


Figure 2. Trace plot for pressure sensor from a single wafer run.

3.2. Modelling

Once appropriate UVA models are identified and summary statistics calculated, the first pass of modelling results are studied. This typically involves filtering out redundant variables to reduce the dimensions to a more manageable subset. Depending on the problem being analysed, a variety of supervised and unsupervised learning algorithms are available for use. Supervised algorithms, like some regression or classification techniques, help establish relationships between a dependent variable (e.g. metal film thickness) and a set of independent variables (e.g. gas flows); unsupervised methods like Principle Components Analysis (PCA) or clustering help highlight interaction between different variables (sensors).

The first pass modelling results are followed by a number of iterations to successively and systematically remove noise and unwanted variables to improve model quality. Next, a model quality report is generated with the top ranked variables and their respective contributions (Figure 3). A glance at the plot of predicted values vs. actual values (Figure 4) provides a good measure of the model quality.

More often, a combination of two algorithms is significantly more effective than using any one algorithm. For example, consider fitting a prediction model for transistor current (Idrive) after an etching process. It was found that Rules Ensemble alone performed very poorly in terms of the predictive ability of the model. Likewise, an algorithm like Random Forest, which builds decision trees based on splitting sensor values, is prone to overfitting. However, a combination of the two methods proved to be significantly more powerful – a Rules Ensemble algorithm was used to reduce the number of variables (sensors), while Random Forest was then used to build a model with high predictive power and good generalizability.

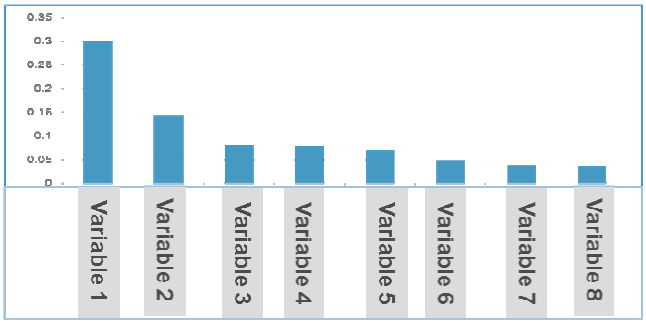


Figure 3: Contribution Pareto of top-ranked variables.

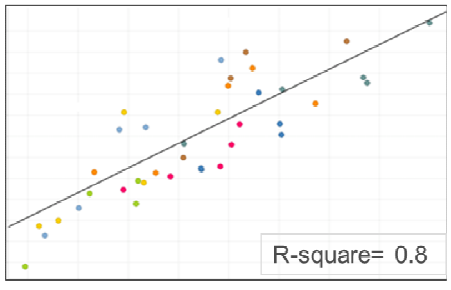


Figure 4: Plot of predicted value of dependent variable (x) vs. actual value of dependent variable (y).

Yield driven control limits were set at the end of the data mining phase wherein the independent variable is allowed to vary in a specific range based on its correlation with the yield numbers as

depicted in Figure 5. Idrive is the dependent variable and the FDC (Fault Detection and Classification) sensor is the independent variable like gas flow that was found to be highly correlated with the dependent variable. The vertical lines on the right side of the chart indicate the independent variable control limits needed to meet the dependent variable specification limits as shown in the left side of the chart.

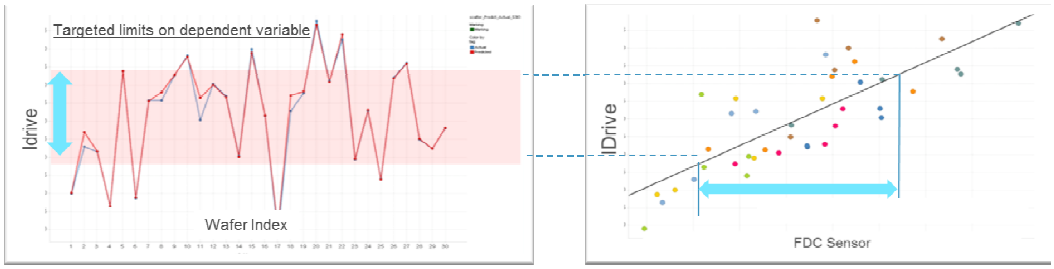


Figure 5. Left: Actual (blue) vs. predicted (red) dependent variable and Right: Scatter plot of independent variable vs. dependent variable.

Figure 6 is the visual representation of the said data mining methodology.

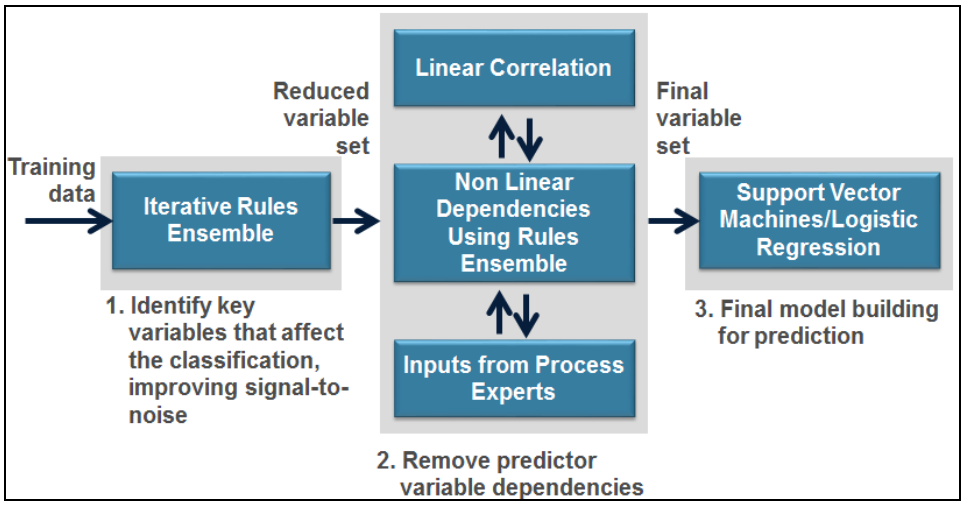


Fig 6. Advanced data mining approach.

4. RESULTS AND DISCUSSION

Some of the high level data mining problem statements that we address are listed below:

- Transistor drive current not matched post-etch process.
- Thickness uniformity of hard mask layers are not in specification limits of <1.5%.
- Improving the temperature matching on Epi chambers from ±10ºC to ≤5ºC.
- Identifying key sensors controlling transistor drive current from Epi process.
- Determining the root cause of arc and deep scratches at Copper Chemical and Mechanical Planarization process and reducing them.
- Reducing particle count on various key device layers.

Multivariate regression was performed using Random Forest and Rules Ensemble models separately in R platform in order to find the key sensors controlling the transistor drive current variation across multiple chambers from an etch process. The model prediction power (based on R-square value) was 0.95. From the drill down and physical verification of the model the root cause for transistor drive current variation was identified. Subsequent recipe optimization resulted in reducing the standard deviation of the process by 30% as depicted in Figure 7.

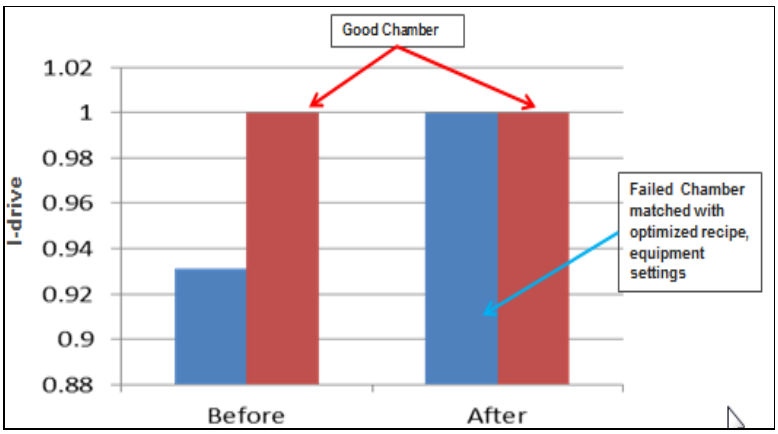


Figure 7. Chamber matching performance matching after changes implemented.

In another case multivariate regression analysis found the root cause sensors affecting the uniformity mismatch of dielectric layers wherein the slope of liquid gas flow sensor across 6 chambers were found to be the variable of highest correlation to the dependent variable (i.e., uniformity). Figure 8 shows the improvement in non-uniformity values after the changes were implemented based on the data mining results on data from 300 wafers.

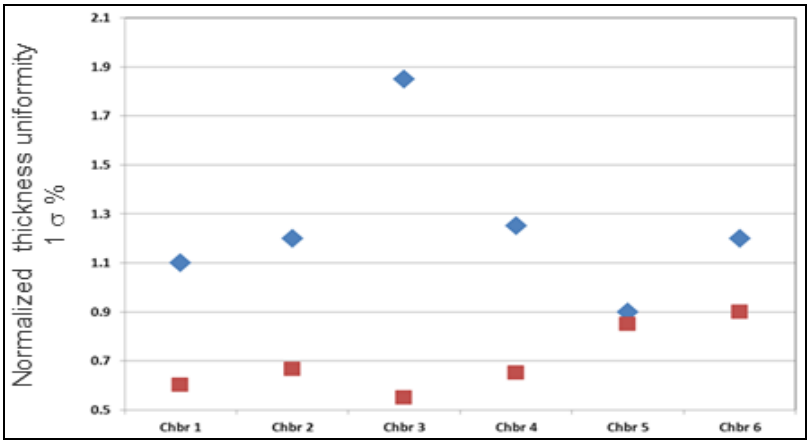


Figure 8. Dielectric film non-uniformity improvement in 6 chambers.

The chart below (Figure 9) shows the pre- and post-implementation results obtained after implementing changes based on the FabVantage data mining analysis. Analysis of gate critical dimension (CD) data over a span of six months revealed that a bad RF generator was causing a bias impedance mismatch (resistance and reactance), which was in turn driving the variation in gate CD.

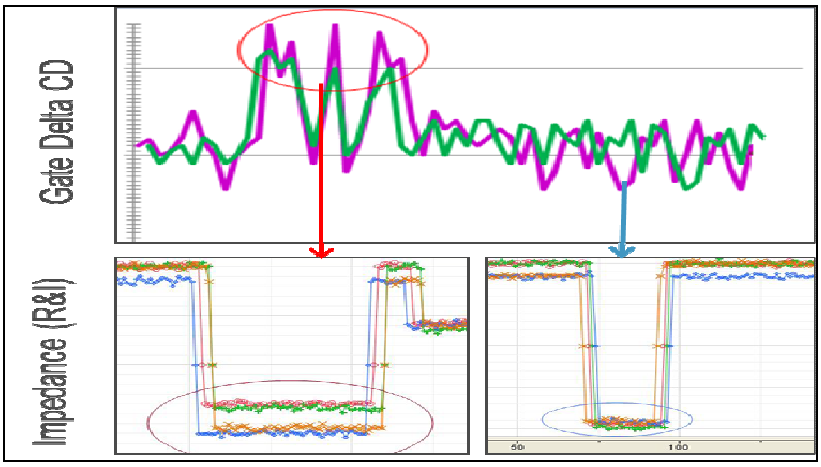


Figure 9. Identification of root cause driving variation in gate CD across 2 etch chambers.

5. CONCLUSIONS

Advanced data mining techniques are deployed in order to achieve higher yields in IC manufacturing units. Different statistical modelling techniques were used to study the impact of independent physical variables on the chamber matching, fault detection and to set the new control limits in order to maintain the yield at desired levels. Similar methodology can be used in any semiconductor, electronic, and photovoltaic manufacturing.

ACKNOWLEDGEMENTS

The authors would like to thank the Applied Fabvantage™ team personnel for their direct or indirect support.

REFERENCES

[1] R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org/.

[2] JMP®, Version 11. SAS Institute Inc., Cary, NC, 1989-2007.

[3] J.H. Friedman, and B. E. Popescu, "Predictive Learning via Rules Ensemble," Stanford University, Department of Statistics, Technical Report, 2005.

[4] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

AUTHORS

Sidda Reddy Kurakula obtained a Master's Degree in Semiconductor Physics in 2007 from Indian Institute of Science, India. He joined Applied Materials in 2008, where he has worked as a process engineer and data analyst. Mr. Kurakula specializes in data mining and analysis of sensor and event data from process equipment. He has performed data analysis for nearly all of Applied Materials' semiconductor and solar processing equipment. His work has helped to identify root causes of tool excursions and mis-processing on many types of equipment, including CVD, Etch, CMP, and implant. Mr. Kurakula has co-authored 2 papers and has contributed to many internal white papers and project reports.

Lokesh Kulkarni received both his Bachelor's and Master's degrees in industrial engineering from Texas A&M University at College Station, Texas (USA). In 2013, Lokesh joined Applied Materials in the capacity of an Engineer – Technology, with much of his work focused on leveraging data mining techniques in the context of semiconductor device manufacturing.

Madhu Babu Dasari obtained a Bachelor's Degree in Computer science from Jawaharlal Nehru University, India. He joined Applied Materials in 2005, where he started working as a software engineer – developer on various modules related to Applied Materials E3 software.

Helen Armer received a Bachelor's Degree in Chemical Engineering from Tulane University in Louisiana (USA), and a Ph.D. in Chemical Engineering from The University of Houston in Texas (USA). She joined Applied Materials in 1999. She has held various roles, including engineering, product management, and knowledge base management. She is currently responsible for the knowledge base for Applied's consulting and advanced services businesses. She holds 26 U.S. patents and has published >25 papers.

INTENTIONAL BLANK

AN EFFECTIVE TOKENIZATION ALGORITHM FOR INFORMATION RETRIEVAL SYSTEMS

Vikram Singh and Balwinder Saini

Department of Computer Engineering,
National Institute of Technology, Kurukshetra, Haryana, India
viks@nitkkar.ac.in
me7saini@gmail.com

ABSTRACT

In the web, amount of operational data has been increasing exponentially from past few decades, the expectations of data-user is changing proportionally as well. The data-user expects more deep, exact, and detailed results. Retrieval of relevant results is always affected by the pattern, how they are stored/ indexed. There are various techniques are designed to indexed the documents, which is done on the token's identified with in documents. Tokenization process, primarily effective is to identifying the token and their count. In this paper, we have proposed an effective tokenization approach which is based on training vector and result shows that efficiency/ effectiveness of proposed algorithm.Tokenization of a given documents helps to satisfy user's information need more precisely and reduced search sharply, is believed to be a part of information retrieval. Tokenization involves pre-processing of documents and generates its respective tokens which is the basis of these tokens probabilistic IR generate its scoring and gives reduced search space. No of Token generated is the parameters used for result analysis.

KEYWORDS

Information Retrieval (IR), Indexing/Ranking, Stemming, Tokenization.

1. INTRODUCTION

Information retrieval is always attracted immense research interest and huge possibility in field of data mining. An IR model concerning with representation, storage, access and retrieval of data relevant to user's query [1] [2]. Following are some current research trends [3] in the area of IR:

- Information Searching
- Ranking/Indexing of user's query results.
- Elaborating representation and storage of information
- Classification of documents (i.e. Pre-defined groups)
- Clustering of documents (i.e. Automatically creates clusters)

Information retrieval system mainly consists of two phases, storing indexed documents and retrieval of relevant results, as shown in figure 1. Phase 1, mainly focus on the identification of

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 109–119, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4910

tokens, and index the tokens based on some parameters [4]. It is clear, that identification of token is important and critical aspect of IR model. Tokenization is a process of identification of token/topics within input documents and it helps to reduced search with significant degree [5]. The secondary advantage of tokenization in effective use of storage space, as it reduces the storage spaces required to store tokens identified from input documents [14]. In modern age of data/information, when data/information is expanding manifold on every day from its origin, in form of documents, web pages etc, so importance of effective and efficient tokenization algorithm become critical for an IR system. There are various traditional techniques for tokenizations are designed, Porter's algorithm is one of the most prominent tokenization among all such techniques, but this algorithm suffers from accuracies during the identification and efficiency [15]. The enhanced algorithm is also designed to overcome the inaccuracy in token identification, but problem still persists. In this paper, an approach is proposed for of tokenization, in which is token identification is completely based on the documents vectors.

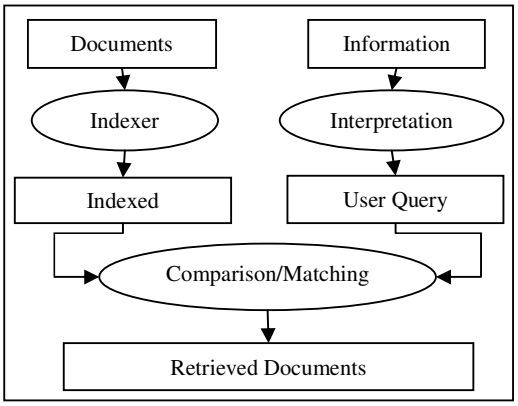


Fig. 1 Formal IR Model System [3]

Tokenization process is an integral part of IR systems, involves pre-processing of given documents and generates respective tokens. In some, tokenization techniques count of token were used to establish a value "Word Count or Token Count" which can be used as indexing/ranking process. A typical structure of tokenization process is explained in figure 2.

Information retrieval models historically many years back to the beginning of written language as information retrieval is related to knowledge stored in textual form [4]. Ranking algorithm/ Indexing algorithm uses the input from tokenization, which is either word count or token count? The affectivity of indexing algorithm is heavily depends upon the quality of token generated by tokenization process.

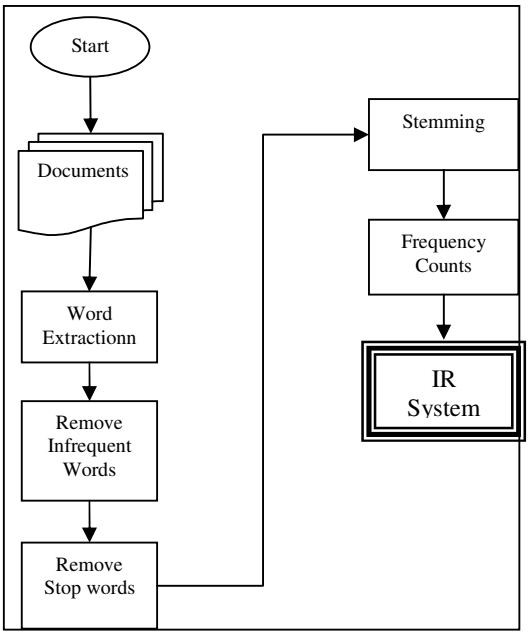


Fig. 2 Tokenization Process

Clearly, the central focus of an IR model is to find the relevant document to issue of finding the relevant document to user's query. Such a decision is usually dependent on an indexing/ranking algorithm which attempts to establish a simple ordering of the documents retrieved [6] [7]. Documents appearing at the top of this ordering are considered to be more likely to be relevant and useful for future patterns. Thus, ranking algorithms are at the core of information retrieval systems.

A ranking algorithm operates according to basic premises (regarding document relevance) yield distinct information retrieval models. The IR model adopted determines the predictions of what is relevant and what is not (i.e. the notion of relevance implemented by the system).

Related Work- Traditional document tokenization techniques are being used in various unsupervised learning approaches for solving problems [7]. Traditional approaches often fail to obtain good tokenization solution when users want to group documents according to their need [9]. An approach to make an effective pre-Processing steps to save both space and time requirements by using improved Stemming Algorithm [11]. Stemming algorithms are used to transform the words in texts into their grammatical root form [11]. Several algorithms exist with different techniques. This is most widely used porter's stemming algorithm [11]. The other enhanced working model is also proposed, in which inaccuracies encountered during the stemming process has been removed by proposing a solutions [9]. The tokenization involves multiple activities to be performed during the life cycle [13]. There are still a lot of scope of improvement on the accuracy of token identification capability of algorithm & efficiency of approach [11][12].

2. INFORMATION RETRIEVAL

Classical retrieval modeling considers documents as bags of words. This stands for the view of the model as an entity without structure where only the numbers of occurrences of terms are important for determining relevance. Whenever a query is posed to a retrieval system every document is scored with respect to the query [3]. The scores are sorted and then final ranked list is presented to the user. A retrieval model is in charge of producing these scores. In general models for retrieval

do not care about efficiency: they solely focus on understanding a user's information need and the ranking process.

- The user's internal cognitive state or information need is turned into an external expression or query based on a query model.
- Each document is assigned a representation that indicates what the document is about and what topics it covers based on a document model.
- A similarity function can be used to estimate the relevance of a document to the information need based on the document model and on the query model.

Therefore the three classic models in information retrieval are called Boolean, Vector and Probabilistic. In the Boolean model documents and queries are represented as set of index terms. Thus we say that the model is set theoretic. In the vector model documents and queries are represented as vectors in the dimensional space. Thus we say that the model is algebraic. In the probabilistic model the framework for modeling document and query representations is based on probability theory [3] [4]. Thus as the name indicates we say that the model is probabilistic.

3. TOKENIZATION

Tokenization is a critical activity in any information retrieval model, which simply segregates all the words, numbers, and their characters etc. from given document and these identified words, numbers, and other characters are called tokens [7] [8]. Along with token generation this process also evaluates the frequency value of all these tokens present in the input documents.

All the phases of tokenization process are shown in figure 2. Pre-processing involves the set of all documents are gathered and passed to the word extraction phases in which all words are extracted [12]. In next phase all the infrequent words are listed and removed for example remove words having frequency less than two. Intermediate results are passed to the stop word removal phase. In this phase remove those english words which are useless in information retrieval these english words are known as stop words.

For example stop words [2] include "the, as, of, and, or, to etc. this phase is very essential in the tokenization because it has some advantages: It reduces the size of indexing file and it also improve the overall efficiency and make effectiveness.

Next phase in tokenization is stemming [2]. Stemming phase is used to extract the sub-part i.e. called as stem/root of a given word [8][9][11]. For example, the words continue, continuously, continued all can be rooted to the word continue. The main role of stemming is to remove various suffixes as result in the reduction of number of words, to have exactly matching stems, to minimize storage requirement and maximize the efficiency of IR Model. The typical stemming process is illustrated in figure 3.

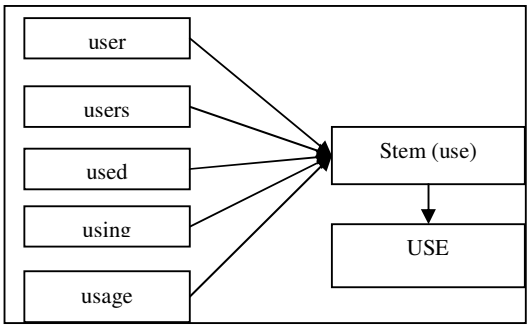


Fig.3 Stemming Process

On the completion of stemming process, next step is to count the frequency of each word. Information retrieval works on the output of this tokenization process for achieving or producing most relevant results to the given users [7] [14].

For example, there is a document in which the information likes “This is an Information retrieval model and it is widely used in the data mining application areas. In our project there is team of 2 members. Our project name is IR”. If this document is passing to the tokenization technique or process then the output of the process is like it separate the words this, is, an, information etc. if there is a number it can also separate from other words or numbers and finally give the tokens with their occurrences count in the given document. This is shown by following example:

Input:

> “This is an Information retrieval model and it is widely used in the data mining application areas. In our project there is team of 2 members. Our project name is IR”.

Fig. 4: Input Document

After applying tokenization process to the figure 4 then the output is formed like:
Output:

Words= this<1> is<4> an<1> Information<1> retrieval<1> model<1> and<1> it<1> widely<1> used<1> in<2> the<1> data<1> mining<1> application<!> areas<1> our<2> project<2> there<1> team<1> of<1> members<1> name<1> IR<1>

Numbers= 2<1>

In angular braces, the value shows the frequency of a word in the given document for example word “our” and “project” occurs two times in the document so their frequency is 2. It also provides the facility to separate the stop words and only gives the distinct words form the given document. In this paper, tokenization process plays a crucial part of finding distinct keywords and their respective frequency values present in the document. The tokenization technique, which tokenize all the documents and then applying the working principle of probabilistic information retrieval model on the output of this tokenization technique for finding their probability scores it extends the overall ranking process for obtaining better results.

4. PROPOSED ALGORITHM AND EXAMPLE

In the proposed algorithm, tokenization is done based on the set of training vectors which are initially provided into the algorithm to train the system. The training documents are of different knowledge domain, are use to create vectors. The created vector helps algorithm to process the input documents. The tokenization on documents is performed with respect to the vectors, use of vectors in pre tokenization helps to make whole tokenization process more precise and successful. The effect on tokenization of vectors is shown in results section also, where the no of token generated & time consumed for the process significantly differ. Following figure-5 shows the proposed algorithm for the tokenization of documents.

Input (Di)
Output (Tokens)
Begin
Step 1:
Collect Input documents (Di) where i=1, 2, 3….n;
Step2:
For each input Di;
Extract Word (EWi) = Di;
// apply extract word process for all documents i=1, 2, 3…n in and extract words//
Step 3:
For each EWi;
Stop Word (SWi) =EWi;
// apply Stop word elimination process to remove all stop words like is, am, to, as, etc. //
Stemming (Si) = SWi;
// It create stems of each word, like "use" is the stem of user, using, usage etc. //
Step 4:
For each Si;
Freq_Count (WCi)= Si;
// for the total no. of occurrences of each Stem Si. //
Return (Si);
Step 5:
Tokens (Si) will be passed to an **IR System**.
End

Example:-**Phase 1:**

Input Documents:

S.No.	Documents Contents
doc1	Military is a good option for a career builder for youngsters. Military is not covering only defense it also includes IT sector and its various forms are Army, Navy, and Air force. It satisfies the sacrifice need of youth for their country.
doc2	Cricket is the most popular game in India. In crocket a player uses a bat to hit the ball and scoring runs. It is played between two teams; the team scoring maximum runs will win the game.

doc3	Science is the essentiality of education, what we are watching with our eyes happening non-happening all include science. Various scientists working on different topics help us to understand the science in our lives. Science is continuous evolutionary study, each day something new is determined.
doc4	Engineering makes the development of any country, engineers are manufacturing beneficial things day by day, as the professional engineers of software develops programs which reduces man work, civil engineers gives their knowledge to construction to form buildings, hospitals etc. Everything can be controlled by computer systems nowadays.

Input four documents to the tokenization process, the process will complete the action in following steps,

Phase 2:

In this phase, all the words are extracted from these four documents as shown below:

Name: doc1

[Military, is, a, good, option, for, a, career, builder, for, youngsters, Military, is, not, covering, only, defense, it, also, includes, IT, sector, and, its, various, forms, are, Army,, Navy,, and, Air, force., It, satisfies, the, sacrifice, need, of, youth, for, their, country.]

Name: doc2

[Cricket, is, the, most, popular, game, in, India., In, crocket, a, player, uses, a, bat, to, hit, the, ball, and, scoring, runs., It, is, played, between, two, teams;, the, team, scoring, maximum, runs, will, win, the, game.]

Name: doc3

[Science, is, the, essentiality, of, education,, what, we, are, watching, with, our, eyes, happening, non-happening, all, include, science., Various, scientists, working, on, different, topics, help, us, to, understand, the, science, in, our, lives., Science, is, continuous, evolutionary, study,, each, day, something, new, is, determined.]

Name: doc4

[Engineering, makes, the, development, of, any, country,, engineers, are, manufacturing, beneficial, things, day, by, day,, as, the, professional, engineers, of, software, develops, programs, which, reduces, man, work,, civil, engineers, gives, their, knowledge, to, construction, to, form, buildings,, hospitals, etc., Everything, can, be, controlled, by, computer, systems, nowadays.]

Phase 3 and Phase 4:

After extracting all the words, next phases is to remove all stop words and stemming, as shown below:

Name: doc1

[militari, good, option, for, career, builder, for, youngster, militari, not, cover, onli, defens, it, also, includ, it, sector, it, variou, form, ar, armi, navi, air, forc, it, satisfi, sacrific, need, youth, for, their, country]

Name: doc2

[cricket, most, popular, game, in, india, in, crocket, player, us, bat, to, hit, ball, score, run, it, plai, between, two, team, team, score, maximum, run, win, game]

Name: doc3

[scienc, essenti, educ, what, we, ar, watch, our, ey, happen, non, happen, all, includ, scienc, variou, scientist, work, on, differ, topic, help, to, understand, scienc, in, our, live, scienc, continu, evolutionari, studi, each, dai, someth, new, determin]

Name: doc4

[engine, make, develop, ani, countri, engin, ar, manufactur, benefici, thing, dai, by, dai, profession, engin, softwar, develop, program, which, reduc, man, work, civil, engin, give, their, knowledg, to, construct, to, form, build, hospit, etc, everyth, can, be, control, by, comput, system, nowadai]

Now, as above mentioned, the documents are ready to process by information retrieval model. All the comparative improvement on the performance in algorithm is discussed in subsequent section.

5. RESULTS AND EXPERIMENTS

In this section, the results are shown, the comparison on both cases tokenization with vectors (with pre-processing) and tokenization without vectors (without pre-processing) on given input documents are shown. The results shown in the paper are of are based on the experimentation over more than 100 input documents and more than 50 input document vectors. Further, for the comparative analysis below mentioned parameters are used:

(1) ***Number of Tokens Generated***: Total no of tokens/topic generated distinctly in one input documents after processing are one of the parameter for result analysis. This number varies in both scenario's, as tokenization with pre-processing generate more accurate and effective token with respect to input document, which results less storage space required and more accurate results to the user. Tokenization without processing leads to large number of tokens, which is difficult to store and affects user results adversely.

(2) ***Strategy***: There are two alternatives of strategy, tokenization with pre-processing and tokenization without pre-processing. Pre-processing involves creation of document vectors based on training documents and then identifying token on input documents based with respect to vectors. The tokenization with pre-processing generates more accurate and effective tokens with more efficient manner, while in without pre-processing strategy simply parses input documents and generates tokens.

(3) ***Overall-Time Value***: Time consumed in entire tokenization process is directly proportional to performance measure of an IR system, as it deeply affects the Indexing & storage aspects.

The performance analysis shown in figure 6 is between strategy (tokenization with pre-processing or without pre-processing) and number of tokens generated. As mentioned previously also, the tokenization with pre-processing generates less no of tokens but the tokens are accurate with in context of result retrieval. In tokenization with pre-processing 200 numbers of tokens generated while for same set of input documents another strategy (without pre-processing) generates more than 300 tokens. The more is the number of token generated, bigger is the challenge to manage them into storage space & effect in accuracy of results retrieval.

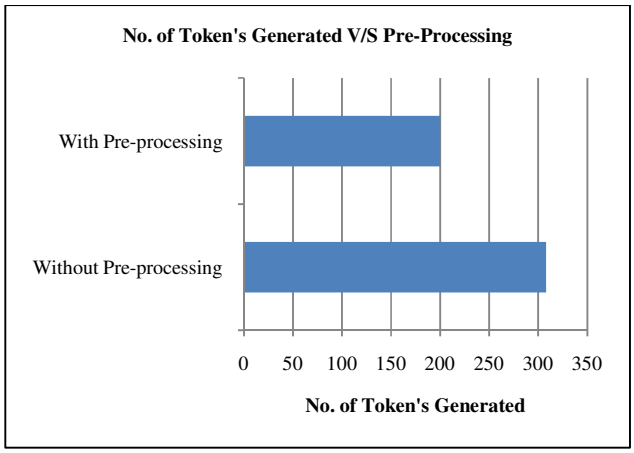


Fig. 6: Document Tokenization Graph

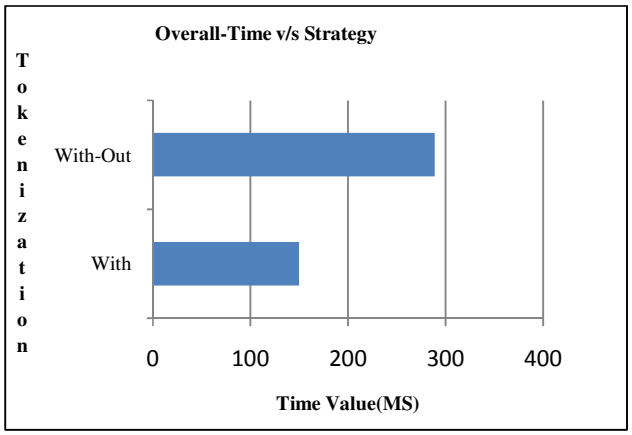


Fig.7: Overall-Time Graph

Another result graph is shown in figure 7, overall time consumed by the strategy is an important factor, which affects overall efficiency of an IR system. The Tokenization with Pre-processing leads to effective and efficient approach of processing, as shown in results strategy with pre-processing process 100 input documents and generate 200 distinct and accurate tokens in 156 (ms), while processing same set of documents in another strategy takes 289 (ms) and generates more than 300 tokens.

6. CONCLUSION

IR model centrally focused on providing relevant results to the user. Relevancies of retrieved results are deeply affected with the quality of indexing / ranking algorithm. Finding information is not the only activity that exists in an Information Retrieval (IR) system. Indexing process, represent into document based on some score like word count, which is generally obtained from tokenization process. There are various traditional techniques for tokenizations are designed, Porter's algorithm is one of the most prominent tokenization among all such techniques, but this algorithm suffers from accuracies during the identification and efficiency. The enhanced algorithm is also designed to overcome the inaccuracy in token identification, but problem still persists. In this paper, an

approach is proposed for of tokenization, in which is token identification is completely based on the documents vectors. The documents vectors are created after the training process. The vectors plays critical role in overall token identification and make entire process effective and efficient. The performance of different information retrieval models are governed by some conditions which are to be outlined. In the results, it shown that tokenization with pre-processing generates better tokens, as it is with less number of token generated and less storage space is required and it facilitates with more accuracy in results retrieval and this is also responsible for reducing the overall-time value of information retrieval model. This algorithm performs better than traditional algorithm of tokenization because of the accuracy in token identification phase.

REFERENCES

[1] G. Salton, M.J. Mcgill, "Introduction to Modern Information Retrieval", Mcgraw-Hill Book Co., New York, 1983.

[2] R. Baeza-Yates, B. Ribeiro –Neto, "Modern Information Retrieval", Harlow: Acm Press 1999.

[3] B. Saini, V. Singh, S. Kumar, "Information Retrieval Models And Searching Methodologies: Survey", In International Journal Of Advance Foundation and Research in Computer, pp. 57-62, 2014.

[4] H. Dong, F.K. Husain, E. Chang, "A Survey in Traditional Information Retrieval Models", IEEE International Conference on Digital Ecosystems and Technologies, Pp. 397-402, 2008.

[5] S. Raman, V. Kumar, S. Venkatesan, "Performance Comparison of Various Information Retrieval Models Used in Search Engines", IEEE Conference on Communication, Information and Computing Technology, Mumbai, India, 2012.

[6] J. Hua, "Study on the Performance of Information Retrieval Models", In 2009 International Symposium on Intelligent Ubiquitous Computing and Education, Pp. 436-439, 2009.

[7] J. Qui, C. Tang, "Topic Oriented Semi-Supervised Document Clustering", In Proceedings of SIGMOD, Workshop on Innovative Database Research, pp- 57-52, 2007.

[8] M. Karthikeyan, P. Aruna, "Probability Based Document Clustering and Image Clustering using Content-Based Image Retrieval", In Elsevier Journal of Applied Soft Computing, Pp.959-966, 2012.

[9] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE TRANSACTIONS ON KNOWLEDGE Fadi Yamout, "Further Enhancement to the Porter's Stemming Algorithm", Issue 2006

[10] V. Srividhya, R. Anitha, "Evaluating Preprocessing Techniques in Text Categorization - International Journal of Computer Science and Application" Issue 2010.

[11] C.Ramasubramanian, R.Ramya, Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 12, December 2013

[12] Karbasi, S, Boughanem, M. (2006) "Document length normalization using effective level of term frequency in large collections, Advances in Information Retrieval, Lecture Notes in Computer Science", Springer Berlin / Heidelberg, Vol. 3936/2006,Pp.72-83.

[13] Diao, Q, Diao, H. (2000) "Three Term Weighting and Classification Algorithms in Text Automatic Classification", The Fourth International Conference on High-Performance Computing in theAsia-Pacific Region, Vol. 2, P.629.

[14] S. K. M. Wong, W. Ziarko, P. C. N. Wong, "Generalized vector space model in information retrieval," in the 8th Annual International ACM SIGIR Conference on Research and Development n Information Retrieval, New York, 1985, pp. 18-25.

[15] Xue, X, Zhou, Z. (2009) "Distributional Features for Text Categorization", IEEE Transactions on Knowledge and Data Engineering, Vol.21, No. 3, Pp. 428-442.

AUTHORS

Vikram Singh (Assistant Professor at NIT Kurukshetra) received his Master of Technology from School of Computer& Systems Sciences, JNU (New Delhi), 2009. He is currently pursuing PhD in Computer Engineering department at National Institute of Technology, Kurukshetra, Haryana, India. His areas of research includes: Distributed Databases System, Query Processing, Ontology and Semantic Web & Big Data Mining.

Balwinder Saini received his bachelor degree in Computer Science and Engineering from Kurukshetra University, Haryana, India, 2011. He is currently pursuing M.Tech degree in Computer Engineering at National Institute of Technology, Kurukshetra, India. His research interests include: Advance Data Base Management System, Knowledge base system and Information Retrieval.

INTENTIONAL BLANK

COLOR IMAGE RETRIEVAL BASED ON FULL RANGE AUTOREGRESSIVE MODEL WITH LOW-LEVEL FEATURES

A. Annamalai Giri[1] and K. Seetharaman[2]

[1]Department of Computer Science, Sri Kuvempu First Grade College , Kengal, Channapatna, Ramanagara Dist, Bangalore, INDIA,
girikumar246@gmail.com
[2]Department of Computer Science & Engineering, Annamalai University, Annamalai Nagar, INDIA,
kseethaddeau@gmail.com

ABSTRACT

This paper proposes a novel method, based on Full Range Autoregressive (FRAR) model with Bayesian approach for color image retrieval. The color image is segmented into various regions according to its structure and nature. The segmented image is modeled to RGB color space. On each region, the model parameters are computed. The model parameters are formed as a feature vector of the image. The Hotlling T^2 Statistic distance is applied to measure the distance between the query and target images. Moreover, the obtained results are compared to that of the existing methods, which reveals that the proposed method outperforms the existing methods.

KEYWORDS

FRAR model, query image, target image, feature vector, spatial features.

1. INTRODUCTION

In computer vision, feature representation schemes are classified as low-level, intermediate, and high-level. The low-level features are represented at pixel level while the high-level features are represented with abstract concepts, and the intermediate-level features represent something in between them. In this paper, it is believed that the low-level features make clear physical meanings and also related to high-level perceptual concepts. Because, the low-level features such as color, texture and spatial orientation of the pixels play vital role in color image formation. The spatial orientation of the pixels forms a shape or structure in an image. Next, the color features refine and enhance the shapes, and perceptually distinguish the regions from each other. Thus, most of the works use low-level features such as color, shape, texture and spatial orientation, and that are used for mine and retrieve the images. One of the most important issues in an image retrieval system is the feature extraction process, where the visual content of the images is mapped into a new space called as feature space. Mainly, the key issue to develop a successful retrieval system relies on identifying and choosing the right features that represent the images as strong as possible. Feature representation of the images may include color [1,2,3,4,5,6,9], texture ([3,7,10,11,13] and shape [1,6,7,8] information.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 121–130, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4911

It is observed from the literature that the low-level features such as texture attributes -- fine texture, coarse texture, texture description, and spatial orientation of the texture play a significant role in image processing, viz. image classification, segmentation, edge and boundary detection, etc. Wu et al. [10] proposed a texture descriptor for browsing and similarity retrieval, and they have reported that the descriptor yield good results. The spatial orientation of the textures is not still used effectively in image retrieval domain. These texture features are more effective than other low-level features. This motivated us to develop the proposed method.

In this paper, a novel technique based on Full Range Autoregressive (FRAR) model is proposed, which characterizes the texture properties such as spatial orientation of the texture and untexture or structure (edge, boundaries) properties. It also gives unique representation to the characterized textures. The FRAR model coefficients K, α, θ and ϕ are considered as features, because α represents the strength of the linear dependency of the pixels on its neighboring pixels; the other parameters θ and ϕ are associated with the circular functions *sine* and *cosine* and their values ranging from 0 to 1; K is a real valued function and it follows t-distribution. Since the parameter, α, represents the strength of the dependency of the pixels, it extracts the textures attributes, i.e. fine, coarse etc. The FVs are compared region wise to that of the feature vectors of the target images in the image database. The distance measure *Hotlling* T^2 *Square* [14] distance is employed to measure the distance between the query and target images.

2. PROPOSED MODEL FOR IMAGE RETRIEVAL

Let $X(k, l)$ be a two-dimentional random variable that represents the intensity value of a pixel at location (k, l) in an image. It is assumed that the pixel, X, may include noise and is considered independent and identical to a distributed Gaussian random variable with discrete time space and continuous state space with mean zero and variance σ^2 and is denoted as $\varepsilon(k, l)$, i.e. $\varepsilon(k,l) \sim N(0,\sigma^2)$. Thus, the images are assumed to be affected by a Gaussian random process.

Thus, we propose an FRAR model as in equation (1),

$$X(k, l) = \sum_{\substack{p=-M \\ p=q\neq 0}}^{M} \sum_{q=-M}^{M} \Gamma_r X(k + p, l + q) + \varepsilon(k, l) \tag{1}$$

$$\text{where, } \Gamma_r = \frac{K \sin(r\theta)\cos(r\phi)}{\alpha^r} \tag{2}$$

and K, α, θ and ϕ are real parameters. The Γ_rs are the model coefficients, which are computed by substituting the model parameters K, α, θ and ϕ in equation (2). The model parameters are interrelated.

The initial assumptions on the parameters are $K \in R$, $\alpha > 1$, and $\theta, \phi \in [0, 2\pi]$. Further restriction on the range of the parameters is placed by examining the identifiability of the model. It is interesting to note that some of the models used in the previous works such as white noise, Markov random field models and autoregressive models with finite and infinite orders can be regarded as a special case of the proposed FRAR model.

Thus,

i) if we set $\theta = 0$, then the FRAR model reduces to the white noise process.

ii) when α is large, the coefficients $\Gamma_r s$ become negligible as r increases. Hence, the FRAR model reduces to a AR(r) model approximately, for a suitable value of r, where r is the order of the model.

iii) when α is chosen to be less than one, the FRAR model becomes an explosive infinite order model.

The fact that $X(k, l)$ has regression on its neighborhood pixels gives rise to the terminology of Markov process. However, in this case, the dependence of $X(k, l)$ on neighborhood values may be true to some extent. In fact, the process is Gaussian under the assumption that the $\varepsilon(k, l)$s are Gaussian, and in this case, its probabilistic structure is completely determined by its second order properties. Second order properties meant for the proposed FRAR model is asymptotically stationary up to order two, provided $1-\alpha < K < \alpha - 1$. Finally, the range of the parameters of the model is set with the constraints $K \in R$, $\alpha > 1$, $0 < \theta < \pi$, $0 < \phi < \pi/2$.

3. PARAMETER ESTIMATION

In order to implement the proposed FRAR model, we have to estimate the parameters. The parameters, K, α, θ and ϕ are estimated by taking the suitable prior information for the hyper parameters β, γ, and δ, based on Bayesian methodology. Only for the computational purpose, the pixel values of each subimage are arranged as one-dimensional vectors $X(t)$, t =1, 2, 3, . . ., N. Since, the error term $\varepsilon(k, l)$ in equation (1) is independent and identical to distributed Gaussian random variable, the joint probability density function of the stochastic process $\{X(t)\}$ is

$$P\left(\frac{X}{\Theta}\right) \propto \left(\sigma^2\right)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}\left\{X_t - K\sum_{r=1}^{\infty} S_r X_{t-r}\right\}^2\right] \tag{3}$$

where $X = (X1, X2, \ldots, XN)$; $\Theta = (K, \alpha, \theta, \phi, \sigma^2)$ and $S_r = \dfrac{\sin(r\theta)\cos(r\phi)}{\alpha^r}$.

When we analyse the real data with finite number of N observations, the range of the index r, viz. 1 to ∞, reduces to 1 to N, and so the joint probability density function of the observations is given in equation (3). The summation $\sum_{r=1}^{\infty}$ can be replaced by $\sum_{r=1}^{N}$ which gives

$$P\left(\frac{X}{\Theta}\right) \propto \left(\sigma^2\right)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}\left\{X_t - K\sum_{r=1}^{N} S_r X_{t-r}\right\}^2\right] \tag{4}$$

By expanding the square in the exponent, we get

$$P\left(\frac{X}{\Theta}\right) \propto \left(\sigma^2\right)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\left\{T_{00} + K^2\sum_{r=1}^{N} S_r^2 T_{rr} + 2K^2\sum_{\substack{r,s=1 \\ r<s}}^{N} S_r S_s T_{rs} - 2K\sum_{r=1}^{N} S_r T_{0r}\right\}\right] \tag{5}$$

where $T_{rs} = \sum_{t=1}^{N} X_{t-r} X_{t-s}$, $r, s = 0, 1, 2, \ldots, N$

The above joint probability density function can be written as

$$P\left(X/\Theta\right) \propto \left(\sigma^2\right)^{-N/2} \exp\left[-\frac{Q}{2\sigma^2}\right] \tag{6}$$

$$\text{where } Q = T_{00} + K^2 \sum_{t=1}^{N} S_r^2 T_{rr} + 2K^2 \sum_{\substack{r,s=1 \\ r<s}}^{N} S_r S_s T_{rs} - 2K \sum_{r=1}^{N} S_r T_{0r} \tag{7}$$

$$K \in R,\ \alpha > 1,\ 0 < \theta < \pi,\ 0 < \phi < \pi/2 \text{ and } \sigma^2 > 0.$$

So, the joint prior density function of Θ is given by

$$P(\Theta) \propto \beta \exp(-\beta(\alpha-1) - \nu/\sigma^2)(\sigma^2)^{-(\delta+1)}; \quad \sigma^2 > 0,\ \alpha > 1,\ 0 < \theta < \pi,\ 0 < \phi < \pi/2. \tag{8}$$

where, P is a general notation for the probability density function of the random variables given within the parentheses following P.

Using (6), (8) and Bayes theorem, the joint posterior density of K, α, θ, ϕ and σ^2 is obtained as

$$P\left(\Theta/X\right) \propto \exp(-\beta(\alpha-1)) \exp(-1/2\sigma^2)(Q+2\nu)\left(\sigma^2\right)^{-\left(\frac{N}{2}+\delta+1\right)}; \tag{9}$$

$$K \in R,\ \alpha > 1,\ 0 < \theta < \pi,\ 0 < \phi < \pi/2 \text{ and } \sigma^2 > 0.$$

Integrating (9) with respect to σ^2, the posterior density of K, α, θ and ϕ is obtained as

$$P(K, \alpha, \theta, \phi / X) \propto \exp(-\beta(\alpha-1))\ (Q+2\nu)^{-\left(\frac{N}{2}+\delta\right)}; \quad K \in R,\ \alpha > 1,\ 0 < \theta < \pi,\ 0 < \phi < \pi/2 \tag{10}$$

where,

$$[Q+2\nu] = \left[\left(K^2 \sum_{r=1}^{N} S_r^2 T_{rr} + 2K^2 \sum_{\substack{r,s=1 \\ r<s}}^{N} S_r S_s T_{rs} - 2K \sum_{r=1}^{N} S_r T_{0r}\right) + T_{00} + 2\nu\right] \tag{11}$$

That is,

$$\begin{aligned}(Q+2\nu) &= aK^2 - 2Kb + T_{00} + 2\nu, \\ &= C\left[1 + a_1\left(K - b_1\right)^2\right]\end{aligned} \tag{12}$$

where,

$$C = T_{00} - \frac{b^2}{a} + 2\nu$$

$$a = \sum_{r=1}^{N} S_r^2 T_{rr} + 2 \sum_{\substack{r,s=1 \\ r<s}}^{N} S_r S_s T_{rs}$$

$$b = \sum_{r=1}^{N} S_r T_{0r} \; ; \; a_1 = \frac{a}{C} \; ; \; b_1 = \frac{b}{a}$$

Thus, the above joint posterior density function of K, α, θ and ϕ can be rewritten as

$$P(K, \alpha, \theta, \varphi / X) \propto \exp(-\beta(\alpha - 1)) \left[C \left\{ 1 + a_1 (K - b_1)^2 \right\} \right]^{-d} \qquad (13)$$

$$K \in R, \; \alpha > 1, 0 < \theta < \pi, \; 0 < \varphi < \pi/2$$

where, $d = \frac{N}{2} + \delta$

This shows that given α, θ and ϕ the conditional distribution of K is 't' distribution located at b1 with (2d–1) degrees of freedom.

The proper Bayesian inference on K, α, θ and ϕ can be obtained from their respective posterior densities. The joint posterior density of α, θ and ϕ, namely, P(α, θ, ϕ / *X*), can be obtained by integrating (13) with respect to K. Thus, the joint posterior density of α, θ and ϕ is obtained as

$$P(\alpha, \theta, \varphi / X) \propto \exp(-\beta(\alpha - 1)) C^{-d} a_1^{-½} \qquad (14)$$

$$\alpha > 1, \; 0 < \theta < \pi, \; 0 < \phi < \pi/2$$

The point estimates of the parameters α, θ and ϕ may be taken as the means of the respective marginal posterior distribution i.e. posterior means. With a view to minimize the computations, we first obtain the posterior mean of α numerically. Then fix α at its posterior mean and evaluate the conditional means of θ and ϕ fixing α at its mean. We fix α, θ and ϕ at their posterior means respectively and then evaluate the conditional mean of K.

Thus, the estimates are

$$\begin{aligned} \hat{\alpha} &= E(\alpha) \\ (\hat{\theta}, \hat{\phi}) &= E(\theta, \phi / \alpha = \hat{\alpha}) \quad \textit{and} \\ \hat{K} &= E(K / \hat{\alpha}, \theta = \hat{\theta}, \phi = \hat{\phi}). \end{aligned} \qquad (15)$$

The estimated parameters K, α, θ and ϕ are adopted to compute the coefficients $\Gamma_r s$ of the model in equation (1). The model parameters are utilised to compute the autocorrelation coefficient, which is used to identify the texture primitives and micro textures present in the image.

The parameters of the model are combined together and formed as FVs. The same kind of features are also extracted from the target image to be retrieved from the image database. These features are treated as two different FVs $\vec{x}$ and $\vec{y}$.

4. SIMILARITY MEASURE

In order find the similarity of the query and target images, the *Hotlling* T^2 *statistic* (T^2) [14] expressed in equation (16) is applied on the feature vectors $\vec{x}$ and $\vec{y}$ of the query and target images.

$$T_d^2(X,Y) = \frac{n_x \cdot n_y}{n_x + n_y}(\overline{X} - \overline{Y})' A^{-1}(\overline{X} - \overline{Y}) \tag{16}$$

Where,

$$\overline{X} = \frac{1}{n_x}\sum_{i=1}^{n_x} X_i \quad , \quad \overline{Y} = \frac{1}{n_y}\sum_{i=1}^{n_y} Y_i$$

are the sample means, and

$$A_x = \frac{\sum_{i=1}^{n_x}(X_i - \overline{X})(X_i - \overline{X})^T}{n_x - 1}$$

$$A_y = \frac{\sum_{i=1}^{n_y}(Y_i - \overline{Y})(Y_i - \overline{Y})^T}{n_y - 1}$$

$$A = \frac{A_x + A_y}{n_x + n_y - 2}$$

is the unbiased pooled covariance matrix of the query and target images.

If $T_d^2 \le \frac{(N_x + N_y - 2)p}{N_x + N_y - p - 1} F_{\alpha;N_x+N_y-p-1}$, then it is inferred that the query and target images are same or similar (i.e., belongs to the same class); otherwise, the two images are different (i.e. belongs to different classes). $F_{\alpha;N_x+N_y-p-1}$ represents values in statistical F-table at the level of significance α with degrees of freedom N_x+N_y-p-1; p represents the size of the feature vectors. Based on the distance values, the images are marked and indexed in ascending order, and the indexed images are retrieved.

5. MEASURE OF PERFORMANCE

In order to validate the performance of the proposed method, the precision and recall measures [12] are used, which are given in equations (17) and (18).

$$\text{Precision (P)} = \frac{\left|\{\text{retrieved images}\} \text{I } \{\text{relevent images}\}\right|}{\left|\text{retrieved images}\right|} \tag{17}$$

$$\text{Recall}(R) = \frac{\left|\{\text{retrieved images}\} I \{\text{relevant images}\}\right|}{\left|\text{relevant Images}\right|} \quad (18)$$

Where $|\cdot|$ returns the size of the set. The precision (P) represents the ratio of the number of images relevant to the query image among retrieved images to the number of retrieved images. The recall (R) represents the ratio of the number of images relevant to the query image among retrieved images to the number of images relevant to the query image.

6. EXPERIMENTS AND RESULTS

In order to implement the proposed method, 1277 color images of size 512×512 pixels have been collected from various sources such as 293 texture images from Brodatz Album; 488 images from Corel image database; 496 images from VisTex image database; and 268 images with size 128×128 are photographed by a digital camera; 257 images with size 128×128 have been downloaded from various websites. To examine the proposed system is invariant for rotation and scaling; the images are rotated through 90^o, 180^o and 270^o degrees, and scaled. Based on this image collection, an image database and their FV database are constructed.

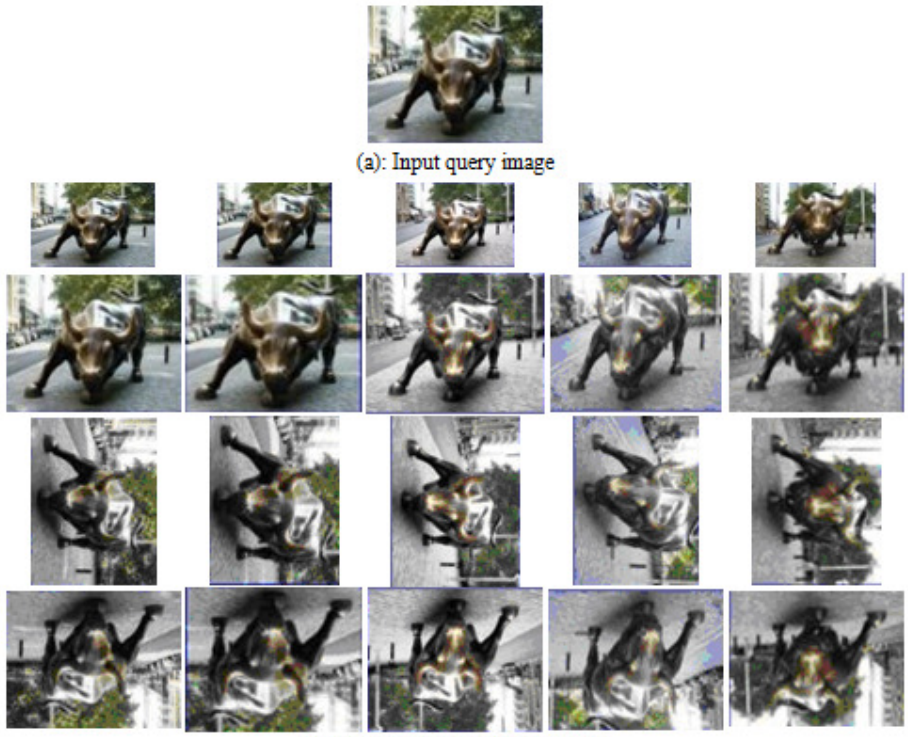


Figure 1. Wall Street Bull – downloaded from internet. (a): Input query image; (b): Retrieved output images: row 1 – scaled down image of size 75 × 100; row 2 – actual image with size 96 × 128; row 3 - images in row 2 are rotated clockwise by 90 degrees; row 4 – images in row 2 are rotated clockwise by 180 degrees.

The input query image is segmented into various regions according to its shapes and structure, and it is modelled to RGB colour space. The proposed method is employed on R, G and B components individually for extracting the image features as discussed in sections 2 and 3. On each region, the parameters and autocorrelation coefficients are computed, and they are combined

together and formed as FVs database. The extracted features are classified into various groups according to their nature using fuzzy c-means algorithm [13]. For each group, median value is calculated, and based on it the FVs are indexed. Based on the classes of the FVs, images in the database are classified into different groups, and it establishes a link between the images and the corresponding FVs of each class. Now, the extracted FVs of the query image are compared to that of the index of the FVs in the image feature database and is identified using the expression given in equation (15). Then the FVs of the query image are matched with the FV database, and retrieve the same or similar images from that class. If the FVs of the query image do not match with any classes of the FV database, then it is formed as a new FV class.

If the given input image is structured, it is segregated into various regions according to its shapes and structure; if it is texture image, then it considered as it is for retrieving the same or similar images from the image database. The input query image is identified whether it is texture or structure by computing the coefficient of variation (CV), and the CV value is compared to a threshold value t. If CV $> t$ then it is assumed that the input image is structured image, and if CV $< t$ then the input image is assumed to be texture image. The threshold t is fixed as 25%.

(a): Input query image

(b): Retrieved output images

Figure 2: VisTex image database – Structural Texture Images: 1(a) - Input query image; 1(b) - row 1: scaled down image; row 2: actual images; row 3: actual images rotated clockwise by 90 degree; row 4: actual images rotated clockwise by 180 degrees; row 5: actual images rotated clockwise by 270 degrees

In order to validate the proposed system, the image in Figure 1(a) is given as input query image to the system, and it retrieves the images in columns 1, 2 and 3 of Figure 1(b) while the level of significance is fixed at 0.001; the images in column 1, 2, 3 and 4 are retrieved at 0.02 level of

significance; at 0.01significance level, the system retrieves the images in column 1, 2, 3, 4 and 7; at the level of significance 0.1, the system retrieves all the images presented in Figure 1(b).

To emphasis the effectiveness and efficiency of the proposed system, another type of structural texture image with stochastic pattern is given as input, which is presented in Figure 2(a). The proposed system retrieves the images in columns 3 of the Figure 2(b) when the level of significance is fixed at 0.001; the images in column 2, 3 and 4 are retrieved at 0.04 level of significance; at 0.08 level of significance, the system retrieves all the images in Figure 2(b).

7. DISCUSSION AND CONCLUSION

In this paper, a novel scheme is proposed for both structure and texture color image retrieval based on FRAR model with Bayesian approach. The model coefficients are computed using the circular functions *sine* and *cosine* as discussed in section 2, the proposed scheme characterizes the structure and texture primitives in the periodic texture patterns also. Since the FRAR model characterizes the texture primitives and provides unique decimal number, it matches exact images in the image database and retrieves. Because, the proposed system facilitates the user to fix the level of significance for the test statistic T_d^2, the user can fix the significance level α at a desired level by which the user can retrieves only the required same or similar images, and not all the relevant images in the database as in the other existing systems. For example, at or below the significance level 0.001 (1%), the system retrieves only the same image, and the rotated and scaled images of the same query image. But there is one disadvantage that if the same image is in the database, it retrieves the image; otherwise, it results that no image is matching with the query image. This problem can be overcome by fixing the significance level $\boldsymbol{\alpha}$ at more than 0.001, by which the system retrieves the similar images. The user can fix the significance level himself at his convenient. Because the proposed system is a distributional approach, it is also invariant for rotation and scaling.

REFERENCES

[1] Qiu, G and Lam, K-M, (2003). "Frequency Layered Color Indexing for Content-Based Image Retrieval", IEEE Transactions on Image Processing, Vol. 12, No. 1, pp. 102- 113.

[2] Xie, Y., Lu, H. and Yang, M-H., (2013) "Bayesian Saliency via Low and Mid-Level Cues", IEEE Transactions on Image Processing, Vol. 22, No 5, pp. 1689-1698.

[3] Zujovic, J., Pappas, T. N. and Neuhoff, D. L., (2013) "Structural Texture Similarity Metrics for Image Analysis and Retrieval", IEEE Transactions on Image Processing, Vol. 22, No. 7, pp. 2545-2558.

[4] Y. D. Chun, N. C. Kim, I. H. Jang, (2008), "Content-based image retrieval using multi-resolution color and texture features", IEEE Trabsaction on Image Processing, Vol. 10, No.6, pp. 1073-1084.

[5] Mustafa Ozden and Ediz Polat, (2007) "A color image segmentation approach for content-based image retrieval, "Pattern Recognition, Vol. 40, pp. 1318-1325.

[6] Yung-Kuan Chan, Yu-An Ho, Yi-Tung Liu and Rung-Ching Chen, (2008) "A ROI image retrieval method based on CVAAO," Image and Vision Computing. Vol. 26, pp. 1540-1549.

[7] B. M. Mehtre, M. S. Kankanhalli, W. F. Lee, (1997a), "Shape measures for content based image retrieval: A comparison," Information processing and Management, Vol. 33, pp. 319-337.

[8] B. M. Mehtre, M. S. Kankanhalli, W. F. Lee, (1997b) "Content-based image retrieval using a composite color-shape approach, "Information processing and Management, Vol. 34, pp. 109-120.

[9] X. Wan, C. C. J. Kuo, (1998), "New approach to image retrieval with hierarchial color clustering, "IEEE Transactions on Circuits and Systems for Video Technology, Vol. 8, pp. 628-643.

[10] P. Wu, B. S. Manjunath, S. Newsam, H. D. Shin, (2000) "A texture descriptor for browsing and similarity retrieval, "Signal Processing.,Vol. 16, pp. 33-43.

[11] N. Nikolaou, N. Papamarkos, (2002) "Color image retrieval using a fractal signature extraction technique, "Engineering Applications of Artificial Intelligence, Vol. 15, pp. 81-96.

[12] M. Kokare, B.N. Chatterji, P.K. Biswas, (2003) "Comparison of similarity metrics for texture image retrieval, " Proc. IEEE Transactions on TENCON 2003, Conference on Convergent Technologies for Asia-Pacific Region, Vol. 2, pp. 571-575.

[13] K. Seetharaman and N. Palanivel, (2013) "Texture characterization, representation, description and classification based on a family of full range gaussian markov random field model, "International Journal of Image and Data Fusion, Vol. 4, No. 4, pp. 1-24.

[14] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, third ed., John Wiley & Sons, Inc., 2003.

AUTHORS

A.Annamalai Giri pursued his MCA from University of Madras, Chennai, India in 2000 and M.Tech.(Information Technology) from Sathyabama University, Chennai India in 2007. Currently, am pursuing my Ph.D. research degree in M.S. Uuniversity, India, and working as Associate Professor of Computer Science at Sri Kuvempu First Grade College, Bangalore,India. His research interests include Statistical Model based Image Mining. Pattern Recognition, Digital Image Analysis, and Data Mining.

Dr.K.Seetharaman received his M.Sc. degree in Statistics from Annamalai University, hidambaram, India in 1990 and the M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, India in 1995. He received his Ph.D. degree in Computer Science & Engg. from Annamalai University in 2006.He is currently working as Associate Professor in the Department of Computer Science & Engg., Annamalai University. His research interests include Statistical Pattern Recognition, Scene Analysis, Digital Image Analysis, and Data Mining and Knowledge Discovery. He is also actively involved in professional activities, viz. reviewer in journals: Pattern Recognition, Knowledge-based Systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, He is currently serving as Associate Editor of the International Journal – Aloy Journal of Soft Computing and Applications, and the Editorial Board Member of the International Journal of Information Technology and Management Sciences. He is also the member of the professional societies – IEEE, ACM, IAENG, and Indian Science Congress Association.

COLOR IMAGE RETRIEVAL BASED ON NON-PARAMETRIC STATISTICAL TESTS OF HYPOTHESIS

R. Shekhar[1] and K. Seetharaman[2]

[1]Department of Computer Science & Engg., Research scholar,
Manonamainam Sundaranar University, Tirunelveli-627012, INDIA,
[2]Department of Computer Science & Engineering,
Annamalai University, Annamalai Nagar, INDIA,
kseethaddeau@gmail.com

ABSTRACT

A novel method for color image retrieval, based on statistical non-parametric tests such as two-sample Wald Test for equality of variance and Man-Whitney U test, is proposed in this paper. The proposed method tests the deviation, i.e. distance in terms of variance between the query and target images; if the images pass the test, then it is proceeded to test the spectrum of energy, i.e. distance between the mean values of the two images; otherwise, the test is dropped. If the query and target images pass the tests then it is inferred that the two images belong to the same class, i.e. both the images are same; otherwise, it is assumed that the images belong to different classes, i.e. both images are different. The proposed method is robust for scaling and rotation, since it adjusts itself and treats either the query image or the target image is the sample of other.

KEYWORDS

variance, mean, query image, target image, non-parametric tests..

1. INTRODUCTION

In recent years, a number of researchers have turned their attention to the content-based image retrieval (CBIR) system. In CBIR system, the researchers concentrate on developing low-level global visual features, namely color properties, shape, texture, and spatial relationship etc., which are used as query for the retrieval process [1-4]. The method proposed in [5-10; 10-15] classifies or segments the entire image into various regions according to the objects or structures present in the image, and the region-to-region comparison is made to measure the similarity between two images [3, 11,12]. In a region-based system, the user has to provide one or more regions from the query image to start a query session. Automatic and precise extraction of image objects is still beyond the ability of the retrieval system available with the computer vision [13]. Therefore, the above system tends to partition one object into several regions, but none of them is representative of the semantic object.

Content-based image retrieval system gives results with low accuracy and slow response time, because there is a big gap between semantic concepts and low-level image features [14]. A

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 131–138, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4912

concept, relevance feedback, has been developed to bridge the gap [15-18]. In [14], a new relevance feedback approach is proposed, which uses Bayesian classifier and treats positive and negative feedback images with different strategies. In relevance feedback method, the user has to provide positive and negative feedback images to improve the performance of the system. Minka and Picard [18] propose the FourEyes system, which has two disadvantages: (i) it uses the region-to-region similarity measure; (ii) the re-clustering of all the features when a new image is added. Thus, it is not very scalable [5]. To overcome this, Jing et al. [5] proposed a system with the features: (i) it computes probabilistic interpretation and it is used in region matching; (ii) region codebook is used; (iii) the SVM based classifier and clustering techniques are adopted, but it requires high computational effort. Above all these, it requires positive and negative query image examples.

Theoharatos et al. [23] proposed a system, based on multivariate non-parametric test, namely Wald-Wolfowitz test (WW-test), and graph theoretic framework of minimal-spanning-tree (MST). In this work, first, the MST is constructed based on the sample identities of the points taken from the images. Based on the consecutive sequence of identical sample identities, runs of the sample points are computed and the WW-test is employed to identify whether the query and target images are same or not. In this work, the drawbacks are

(i) Construction of the MST demands computational overhead.
(ii) Based on the sample identities of the points, run length of each sample identical identities is computed and then the WW-test is used to identify whether the query and target images are same or not.

In this paper, a unified technique is proposed for automatic image retrieval, based on non-parametric tests such as two-sample Wald Test for equality of variance and Man-Whitney U test. In the proposed technique, mean and variance (first and second moments of the sample points) are used as representatives of both query and target images. The methods proposed in [19, 20] retrieve only the texture images with intensity values ranging from 0-255, i.e. gray-scale images. This motivated us to develop a new method which retrieves color images; both texture and structure images; and invariant for rotation and scaling.

2. PROPOSED TEST STATISTIC FOR SIMILARITY OF IMAGES

Let X be a random variable that represents the intensity value with additive noise of a pixel at location (k, l) in a color image. The pixel $X(k,l) \in \Re^3$ is a linear combination of three colors such as red, green and blue, i.e. $X(k, l) = [r(k, l), g(k, l), b(k, l)]^T$, where T represents the transformation of the vector.

2.1. TEST FOR VARIATION BETWEEN THE QUERY AND TARGET IMAGES

Either the query image or the target image is treated as sample while the other is treated as population. To test whether the two images are same or not, first, the variation among the intensity values of the two images are tested, i.e. $H_0 : \sigma_q = \sigma_t$, where σ_q and σ_t represent the variation among the intensity values of the query and target images respectively. To achieve this, the test for homogeneity of variances is employed, i.e. two-sample Wald Test for equality of variance [22]. The *R* test is nearly as robust as Levene's test and nearly as powerful as the *F* test.
Hypotheses:

$H_0 : \sigma_q = \sigma_t$ (Similarity – query and target images belong to the same class)

$H_a : \sigma_q \neq \sigma_t$ (Non-similarity – query and target images belong to different class)

Test Statistic: The two-sample Wald test statistic [22] defined in equation (1) is applied to test whetheher the images are same or not.

$$R = \frac{\left(S_q^2 - S_t^2\right)^2}{\left(\left(m_{q4} - S_q^4\right)/n_q + \left(m_{t4} - S_t^4\right)/n_t\right)} \quad (1)$$

where, $S_i^4 = (m_4 - m_t^2)/n$; S^2 is the unbiased sample variance; m_{i4} are the fourth central sample moments for the i-th sample and i = q, t; n_q and n_t are the number of pixels in the query and target images respectively.

The query and target images are judged to be same, if $R < c_\alpha$, where, c_α is the point for which the Chi-square distribution has weight α in the right hand tail, then the *R* test rejects H_o at approximately $100\alpha\%$ level; otherwise, the two images are assumed to be different.

2.2 . TEST STATISTIC FOR EQUALITY OF SPECTRUM OF ENERGY BETWEEN THE IMAGES

As discussed in the previous section, if the variation among the intensity values of the query and target images passed the test for homogeneity of variances, then it is proceeded to test the equality of means of the two images. To achieve this, the test for equality of means, i.e. Mann-Whitney U (MWU) test [23] is employed. In general, the actual pixel values either in the query image or the target image may exeed 20, thus value of *U* approaches Gaussian distribution, and thus the null hypothesis can be tested by Z test. The MWU test is a greater powerful than the t-test if the populations are non-normal distributions, i.e. a mixture of normal distributions, and it is also nearly as efficient as the t-test in the case of Gaussian distributions.

The test of hypothesis are assumed as follows.

Hypotheses:

H_0 : The query and target images are belonging to the same class.

H_a : The query and target images are belonging to different class.

The test statistic (Z) is defined as in equation (2),

$$Z = \frac{U - E(U)}{SD_{UCorr}} \quad (2)$$

where,

$$U = \left(n_q \times n_t\right) + \frac{n_q \times \left(n_q + 1\right)}{2} - T_q$$

T_q is the larger of the sum of ranks of either the query image or the target image; n_q and n_t are the number of pixels in the query and target images respectively.

$$E(U) = \frac{n_q \times n_t}{2}$$

$SD_{UCorr} = \sqrt{\frac{n_q \times n_t}{n(n-1)}\left(\frac{n^3 - n}{12} - \sum_{i=1}^{k}\frac{t_i^3 - t_i}{12}\right)}$, and *k* is number of tied ranks; t_i is the number of subjects sharing rank *i*.

$$n = n_q + n_t$$

$$U_q = (n_q \times n_t) \times \frac{n_q \times (n_q + 1)}{2} - T_q$$

$$U_t = (n_q \times n_t) - U_q$$

$$n = n_q + n_t$$

Critical region: It is concluded that the query and target images are same, if $Z < Z_{C_\alpha}$, where Z_{C_α} is the critical value at the level of significance α; otherwise, it is concluded that the images are different**.**

3. IMAGE FEATURE DATABASE CONSTRUCTION AND INDEXING

An image database is constructed using various types of images collected from standard Brodatz album, Vistex, and Corel image databases, and from other sources such as internet and images captured by digital camera.

The feature vectors of the query image are matched with the features in the feature vector database using the test statistic discussed in the previous section. The target images are selected from image database based on the significance level (α) at 20%. The significance level can be fixed by the user at 1% or 5% or 10% or 15% or 20% or at any other level according to the user's requirements. The selected images are indexed (ranked) according to the test statistic values, and the distance value between the query and target images from lowest to highest, i.e. in ascending order. The user can select the top most images from the indexed list according to his requirements.

Measure of Performance

In order to measure the performance of the proposed method, the precision and recall measures [25] are used, which are given in equations (3) and (4).

$$Precision = \frac{|\{Relevant\ Images\} I \{Retrieved\ Images\}|}{|Retrieved\ Images|} \quad (3)$$

$$Recall = \frac{|\{Relevant\ Images\} I \{Retrieved\ Images\}|}{|Relevant\ Images|} \quad (4)$$

4. IMAGE DATABASE DESIGN AND EXPERIMENTAL RESULTS

In order to implement the proposed method, 477 color images of size 512×512 pixels have been collected from various sources, i.e. 152 texture images from Brodatz Album, 176 images from Corel image database and 149 images from VisTex image database. The remaining 58 images with size 128×128 are photographed with digital camera; 43 images with size 128×128 have been

downloaded from internet [30]. The textured images collected from Brodatz, Coral and VisTex image databases are divided into 16 non-overlapping sub-images of size 128×128. To examine the proposed system is invariant for rotation and scaling; the images are rotated by 90^0, 180^0 and 270^0, and scaled.

To validate the proposed system, based on statistical non-parametric tests, the concepts discussed in sections 2.1 and 2.2 are implemented with the image database constructed as discussed above. Due to space constraints, for sample, some of the images considered for the experiment are presented in Figures 1 and 2. The experiment is conducted at various levels of significance for the input query image given in column 1 of Figure 1. The images in columns 2, 3, 4, 5 of Figure 1 are retrieved at level of significance, 0.001; images in columns 6, 7 are retrieved at 0.05 level of significance; at 0.15 level of significance, the system retrieves the images in columns 8, 9, 10 and 11.

Figure 1. Structure images; column 1: input query image; columns 2 – 11: retrieved output images

Furthermore to emphasize the efficiency of the proposed method, the images in column 1 of Figure 2 are given as input query to the system. The images in columns 2, 3, 4, 5 of Figure 1 are retrieved at level of significance, 0.001; images in columns 6, 7 are retrieved with level of significance at 0.03; significance level at 0.12, the system retrieves the images in columns 8, 9, 10 and 11.

The obtained results emphasize that the proposed system is robust for scaling and rotation, since it retrieves the same input query image rotated by 90 degrees, 180 degrees, 270 degrees, and scaled images at the level of significance 0.001.

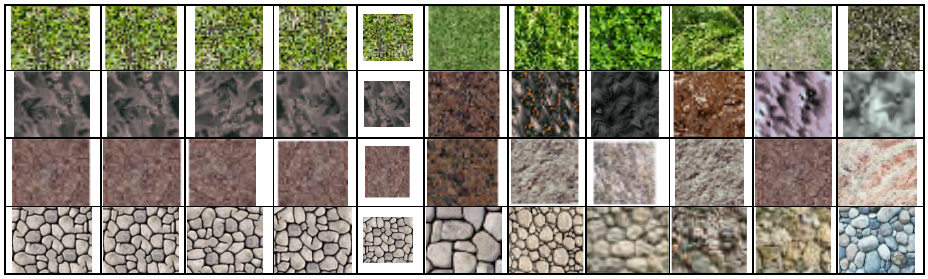


Figure 2. Texture images; column 1: input query image;columns 2 – 11: retrieved output images

Table 1. Performance measure of the proposed method with other existing methods

Database	Proposed method		Statistical Distributional Approach		Orthogonal polynomial		Gabor wavelet		Contourlet Transform	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Brodatz	0.840	0.781	0.836	0.784	0.812	0.685	0.824	0.661	0.825	0.601
VisTex	0.822	0.799	0.810	0.795	0.782	0.681	0.795	0.670	0.805	0.610
Structure Images	0.850	0.824	0.842	0.815	--	--	--	--	0.826	0.621

A comparative study is performed with the existing methods such as Orthogonal polynomial [19], Gabor wavelet transform [20] and the Contourlet transform [27] methods and Statistical distributional approach [28], and the obtained results are presented in Table 1. The results reveal that the proposed method outperforms the existing methods. It is observed from the results that there is no significant difference between the proposed method and the statistical distributional approach. Though both parametric and non-parametric tests yield almost same results; there are some difficulties to employ parametric tests. Since some type of images may not be distributed to Gaussian, at that situation the parametric tests cannot be applied. Thus, the necessity arises to use appropriate non-parametric tests instead of parametric tests.

5. DISCUSSIONS AND CONCLUSION

Block-wise sampling technique proposed in [21] does not yield good results for the rotated and scaled images, because the corresponding blocks of query (actual) and target (transformed) images do not match spatially, while the target image is rotated or scaled. Hence, the technique proposed in [21] fails to match and retrieve the right images. The proposed system avoids this problem, because it uses the global distributional differences of both query and target images; in the case of structured images, these features are extracted from the shapes in both query and target images, and those are compared shape-wise, it compares the number of shapes between the images. The orthogonal polynomial based method [19] retrieve only textured images with gray-scale, and the Gabor features based method [20] retrieves only the textured images in both color and gray-scale. The proposed system retrieves both textured and structured color images, and it is robust for scaled and rotated images. Most of the existing methods retrieve a set of similar images, from which the user has to select the required images. But the proposed system facilitates the user to retrieve the required image only by fixing the level of significance at a desired level.

In this paper, a unified system for both structured and textured color image retrieval is used, based on statistical non-parametric tests of hypothesis, namely test for equality of variances – variation between the query and target images, and the test for equality of means – spectrum of energy. The proposed system is invariant for rotation and scaling, since the query image is treated as either a sample or population of the target image. First, the test for equality of variation between the query and target images is performed; the query and target images pass the test, viz. the two images are same or similar, then the test for equality of mean vectors is performed, i.e. testing the spectrum of energy on the same images. If the query and target images pass these two tests, it is inferred that the two images are identical; otherwise, it is assumed that the images belong to different groups. The proposed system provides hundred percent accuracy and precision, even if either the target or query image is rotated or scaled.

REFERENCES

[1] J.Huang, S.R. Kunar, M.Mitra, W.J. Zhu, and R.Zabih, Image indexing using color correlogram, in: Proc. IEEE Comp. Soc. Conf. Comp. Vis. and Pattern Recognition, vol. 1, 1997, pp. 762-768.

[2] M.Stricker and M. Orengo, Similarity of color images, in: Storage and Retrieval for Image and Video Databases, Proc. SPIE 2420, vol. 1, 1995, pp. 381-392.

[3] A.Pentland, R.Picard, and S. Sclaroff, Photobook: Content-based manipulation of image databases, International Journal of Computer Vision 18(3) (1996) 233-254.

[4] Chiou-Shaan Fuh, Shun-Wen Cho, and Kai Essig, Hierarchical color image region segmentation for content-based image retrieval system, IEEE Transactions on Image Processing 9(1) (2000) 156-162.

[5] F. Jing, M. Li, H.J. Zhang, and B.Zhang, An efficient and effective region-based image retrieval framework, IEEE Transactions on Image Processing 13(5) (2004) 699-709.

[6] Jun-Wei Hsieh and W.Eric L. Grimson, Spatial template extraction for image retrieval by region matching, IEEE Transactions on Image Processing 12(11) (2003) 1404-1415.

[7] S. Belongie, C.Carson, H.Greenspan, and J.Malik, Recognition of images in large databases using color and texture, IEEE Transactions on Pattern Analysis and Machine Intelligence 24(8) (2002) 1026-1038.

[8] F. Jing, B.Zhang, F.Z.Lin, W.Y.Ma, and H.J.Zhang, A novel region-based image retrieval method using relevance feedback, in: Proc. 3rd ACM Int. Workshop on Multimedia Information Retrieval (MIR), 2001.

[9] Y.Deng and B.S.Manjunath, (1999) "An efficient low-dimensional color indexing scheme for region-based image retrieval, in: Proc. IEEE Int. Conf. ASSP", Vol.6, pp. 3017-3020.

[10] Ing-Sheen Hsieh and Huo-Chin Fan, (2001) "Multiple classifiers for color flag and trademark image retrieval", IEEE Transactions on Image Processing, Vol. 10, No.6, pp. 938-950.

[11] J.R. Smith and C.S. Li, (1999), "Image classification and querying using composite region templates, Journal of Computer Vision and Image Understanding", Vol. 75, No. 12, pp. 165-174.

[12] J.Z. Wang and Y.P. Du, (2001) "Scalable integrated region-based image retrieval using IRM and statistical clustering", in: Proc. ACM and IEEE Joint Conference on Digital Libraries, VA, 2001, pp. 268-277.

[13] J.Z. Wang, J. Li, and G.Wiederhold, (2001) "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries", IEEE Transactions of Pattern Analysis and Machine Intelligence, Vol. 23, No. 9, pp. 947-963.

[14] Z. Su, H. Zhang, S.Li, and S. Ma, (2003) "Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning", IEEE Transactions on Image Processing, Vol. 12, No. 8, pp. 924-937.

[15] I.J.Cox, T.P.Minka, T.V.Papathomas, and P.N.Yianilos, (2000) "The Bayesian image retieval system, pichunter: Theory, implementation, and psychophysical experiments", IEEE Transactions on Image Processing, Vol. 9, No. 1, pp. 20-37.

[16] Y.Rui and T.S.Huang, (1999) "Relevance feedback: A power tool for interactive content-based image retrieval", IEEE Circuits Syst. Video Technology, Vol. 8, No. 5, pp. 644-655.

[17] N.Vasconcelos and A.Lippman, (1999) "Learning from user feedback in image retrieval systems", in: Proc. NIPS'99, Denver, CO.

[18] T.P. Minka and R.W. Picard, (1997) "Interactive learning using a society of models", Pattern Recognition, Vol. 30, No. 4, pp. 565-581.

[19] R. Krishnamoorthy, S. Sathiya devi, (2012) "A multiresolution approach for rotation invariant texture image retrieval with orthogonal polynomial model", Journal of Visual Communication and Image Representation, Vol. 23, No. 1, pp. 18-30.

[20] Ju Han, Kai-Kuang Ma, (2007) "Rotation invariant and scale invariant Gabor features for texture image retrieval", Image and Vision Computing, V 25, No. 9, pp. 1474-1481.

[21] Huang, C.W., Lin, K.P., Hung, K.C. (2014) "Intuitionistic fuzzy c-means clustering algorithm with neighborhood attraction in segmenting medical image", Soft Computing, DOI: 10.1007/s00500-014-1264-2.

[22] Allingham, D., Rayner, J. C. W., (2011) "A Nonparametric Two-Sample Wald Test of Equality of Variances", Advances in Decision Sciences, doi:10.1155/2011/748580.

[23] C. Theoharatos, N.A. Laskaris, G. Economou, and S. Fotopoulos (2005) "A Generic scheme for color image retrieval based on the multivariate Wald-Wolfowitz test", IEEE Transactions on Knowledge and Data Engineering, Vol.17, No. 6, pp. 808-820.

[24] M. Kokare, B.N. Chatterji, P.K. Biswas (2003) "Comparison of similarity metrics for texture image retrieval, in: IEEE Proceedings on TENCON 2003 Conference on Convergent Technologies for Asia-Pacific Region, Vol. 2, pp. 571–575.
[25] Powers, David M. W. (2012), "The Problem with Kappa, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL2012) Joint ROBUS-UNSUP Workshop, pp. 345–355.
[26] http://elib.cs.berkley.edu/
[27] Ch. Srinivasa Rao, S. Srinivas kumar, B.N. Chatterji, (2007) "Content based image retrieval using Contourlet Transform", ICGST-GVIP Journal, Vol. 7, No. 3, pp. 9-15.
[28] K. Seetharaman, M. Jeyakarthic, (2014), "Statistical distributional approach for scale and rotation invariant color image retrieval using multivariate parametric tests and orthogonality condition", Journal of Visual Communication Image Representation, Vol. 25, No. 5, pp. 727-739.

AUTHORS

R. Shekhar has vast experience in teaching subjects related to Computer Science & Engineering. He has a post-graduate degree in Computer Networks and is currently pursuing doctoral studies at the MS University; Tirunelveli.He is serving as an Editorial member of IJITCS, an International Journal and has been a reviewer for several National and International Journals. He is Also a Sitting member of different National and international Associations in the field of Computer Science. His Publications appear in many Journals of National and International repute. His researches interests include image processing, Computer Networks and Artificial Intelligence.

Dr. K.Seetharaman received his M.Sc. degree in Statistics from Annamalai University, Chidambaram, India in 1990 and the M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, India in 1995. He received his Ph.D. degree in Computer Science & Engg. from Annamalai University in 2006.He is currently working as Associate Professor in the Department of Computer Science & Engg., Annamalai University. His research interests include Statistical Pattern Recognition, Scene Analysis, Digital Image Analysis, and Data Mining and Knowledge Discovery. He is also actively involved in professional activities, viz. reviewer in journals: Pattern Recognition, Knowledge-based Systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, He is currently serving as Associate Editor of the International Journal – Aloy Journal of Soft Computing and Applications, and the Editorial Board Member of the International Journal of Information Technology and Management Sciences. He is also the member of the professional societies – IEEE, ACM, IAENG, and Indian Science Congress Association.

SKIN COLOUR INFORMATION AND MORPHOLOGY BASED FACE DETECTION TECHNIQUE

M. Sharmila Kumari, Akshay Kumar, Rohan Joe D'Souza, G K Manjunath and Nishan Kotian

Department of Computer Science and Engineering,
PA College of Engineering, Mangalore.
{sharmilabp@gmail.com, akshay.holmes@gmail.com,
dsouza.rohanjoe@gmail.com, manjugreat.gk@gmail.com,
nishank.kotian@gmail.com}

ABSTRACT

Locating and tracking human faces is a prerequisite for face recognition and/or facial expressions analysis, although it is often assumed that a normalized face image is available. In this paper, we propose a faster, yet efficient face detection approach based on mathematical morphology and skin colour information. We have devised some simple post-processing rules to eliminate non-face regions from face regions. Experimentation on our created database is conducted to reveal the performance of the proposed approach.

KEYWORDS

Skin colour, Morphology, Segmentation, Face detection.

1. INTRODUCTION

Human face detection is currently an active research area in the computer vision community. Face localization and detection is often the first step in applications such as video surveillance, human computer interface, and face recognition and image database management. Locating and tracking human faces is a prerequisite for face recognition and/or facial expressions analysis, although it is often assumed that a normalized face image is available. In order to locate a human face, the system needs to capture an image using a camera and a frame-grabber to process the image, search the image for important features and then use these features to determine the location of the face. For detecting face there are various algorithms including skin colour based algorithms. Using skin-colour as a feature for tracking a face has several advantages. Colour processing is much faster than processing other facial features. Under certain lighting conditions, colour is orientation invariant. This property makes motion estimation much easier because only a translation model is needed for motion estimation. However, colour is not a physical phenomenon; it is a perceptual phenomenon that is related to the spectral characteristics of electromagnetic radiation in the visible wavelengths striking the retina.

The accurate detection of human faces in arbitrary scenes is the most important process involved prior to face recognition. When faces could be located exactly in any scene, the recognition step

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 139–147, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4913

afterwards would not be so complicated. Face detection in completely unconstrained settings remains a very challenging task, particularly due to the significant pose and lighting variations. The modern face detectors are mostly appearance-based methods, which mean that they need training data to learn classifiers. Collecting a large amount of ground truth data remains a very expensive task, which certainly demands more research. In environments which have low variations, adaptation could bring very significant improvements to face detection. In this context, we proposed a simple, yet efficient algorithm for face detection which works better even in complex background. The details of the proposed methodology are presented in the following sections. We have created our own database and test the system against the ground truth data.

2. LITERATURE SURVEY

The problem of face detection goes back to early 70's, at that time the overall focus was on finding ways to detect human faces in simple constitutions where typically the face is in a passport like photo with a uniform background and uniform lighting conditions. The research on the subject was rather simple at the time and only came into attention at early 90's where more in-depth research was taking over into the problem using different algorithms, the problem began to address the issues related to detecting faces in complex backgrounds with different scales and rotation degrees, introducing statistical methods and neural networks for face detection in cluttered scenes.

There have been hundreds of reported approaches to face detection. Early Works (before year 2000) had been nicely surveyed in [21] and [4]. For instance, Yang et al. [21] grouped the various methods into four categories: knowledge-based methods, feature invariant approaches, template matching methods, and appearance-based methods. The field of face detection has made significant progress in the past decade. In particular, the seminal work by Viola and Jones [17] has made face detection practically feasible in real world applications such as digital cameras and photo organization software. Mita et al. [10] proposed joint Haar-like features, which is based on co-occurrence of multiple Haar-like features. The authors claimed that feature co-occurrence can better capture the characteristics of human faces, making it possible to construct a more powerful classifier. Another well-known feature set robust to illumination variations is the local binary patterns (LBP) [11], which have been very effective for face recognition tasks [1, 22]. In [5, 23], LBP was applied for face detection tasks under a Bayesian and a boosting framework, respectively. More recently, inspired by LBP, Yan et al. [20] proposed locally assembled binary feature, which showed great performance on standard face detection data sets.

Another popular complex feature for face/object detection is based on regional statistics such as histograms. Levi and Weiss [8] proposed local edge orientation histograms, which compute the histogram of edges orientations in sub-regions of the test windows. These features are then selected by an AdaBoost algorithm to build the detector. Later, Dalal and Triggs [2] proposed a similar scheme called histogram of oriented gradients (HoG), which became a very popular feature for human/pedestrian detection [24, 16, 7, 3].

Training a face detector is a very time-consuming task. In early works, due to the limited computing resources, it could easily take months and lots of manual tuning to train a high quality face detector. A number of papers have been published to speed up the feature process. Wu et al. [18] proposed a cascade learning algorithm based on forward feature selection. Pham and Cham [12] presented another fast method to train and select Haar features. It treated the training examples as high dimensional random vectors, and kept the first and second order statistics to build classifiers from features.

Multiview face detection has also been explored with SVM based classifiers. Li et al. [9] proposed a multiview face detector similar to the approach in [15, 6]. They first constructed a face pose estimator using support vector regression (SVR), and then trained separate face detectors for each face pose. Yan et al. [19] instead executed multiple SVMs first, and then applied an SVR to fuse the results and generate the face pose. This method is slower, but it has lower risk of assigning. Neural networks were another popular approach to build a face detector. Early representative methods included the detectors by Rowley et al. [14] and Roth et al. [13].

It shall be observed from the above brief survey that the face detection in completely unconstrained settings remains a very challenging task, particularly due to the significant pose and lighting variations. In addition, collecting a large amount of ground truth data remains a very expensive task, which certainly demands more research. In this context, we have made an attempt to design a simple yet efficient face detector which works in real environment.

3. PROPOSED METHODOLOGY

Our algorithm is designed to overcome some of the limitations present in the existing algorithms. The existing algorithms use template matching as their primary segmentation algorithm. This process is very time consuming and hence it is proposed to detect face regions without template matching. The model that we proposed is a combination of feature-based and view based approaches. Initially the image is divided based on skin colour. This is the most vital stage as the success of the entire algorithm depends on the efficiency of the segmentation. Thus an extensive time has been given to this stage in the entire process of algorithm building. Once the segmentation of the probable skin regions are identified, then regions to be searched for potential faces have been reduced based on mathematical morphology thus improving the efficiency of the algorithm in terms of time. The very next step is removal of noisy pixels in the image. Here the noise is considered to be a part of the background or any non-face region for that matter. Thus an effective solution has been proposed in terms of morphological operations which eliminate noise and also highlight the face region. This is followed by the design of some heuristic rules based on statistical properties of an image to eliminate non-face regions from the probable face regions. The robustness can be measured from the fact that it works very well for multiple faces in complex backgrounds as it does for simple backgrounds too.

EXPERIMENTAL ILLUSTRATION:

Here, we present the proposed methodology with a sample image shown in Fig. 1.

Fig. 1. Sample image containing face and non-face regions.

Stage1: Segmentation based on Skin colour:

In this stage, probable face regions are identified based on skin color. Prior to segmentation of the skin, the given image is converted into a particular color model. In this case we convert into YCbCr color model. Then thresholds are setup to filter out non-skin pixels based on global thresholding. The probable face regions identified based on skin colour information followed by thresholding is shown in Fig. 2.

Fig. 2. Probable skin regions identified due to skin colour based thresholding.

Stage2: Morphology to eliminate noise:

It shall be observed from Fig. 2 that the skin colour segmentation rejects non-skin colours from the input image. However, the resulting image has quite a bit of noise and clutter. A series of morphological operations are performed to remove the noisy pixels in the image and a masked regions are generated which are placed on the input image to yield skin colour regions without noise and clutter.

Since morphological operations work on intensity images, the colour image is converted into a grey scale image. The intensity thresholding is performed to break up dark regions into many smaller regions so that they can be cleaned up by morphological opening. The threshold is set low enough so that it does not chip away parts of a face but only create holes in it. The morphological opening is performed to remove very small objects from the image while preserving the shape and size of larger objects in the image. A disk shaped structuring element of radius 1 is used. The hole filling is done to keep the faces as single connected regions in anticipation of a second much larger morphological opening. Otherwise, the mask image will contain many cavities and holes in the faces. The morphological opening is performed to remove small to medium objects that are well below the size of a face. A disk shaped structuring element of radius 6 is used in this case. The result of applying the mask to the grey scale version of the input image is shown in Fig. 3.

Fig.3 Mask image generated as the output of morphological operations

Step3: Removal of unwanted regions:

The output from the previous stage contains some non-face regions and hence some heuristic rules are designed to eliminate non-face regions which is achieved by converting a gray scale image to a binary image. The statistical properties of each region are used to eliminate non-face regions from face region. The area of regions in the binary image is then calculated. It shall be observed from Fig. 4 that there are there are numerous non-face regions. Few of the non-face regions occupy small areas and hence these regions can be removed by using a threshold.

Fig.4. Face and Non-face regions after binarization

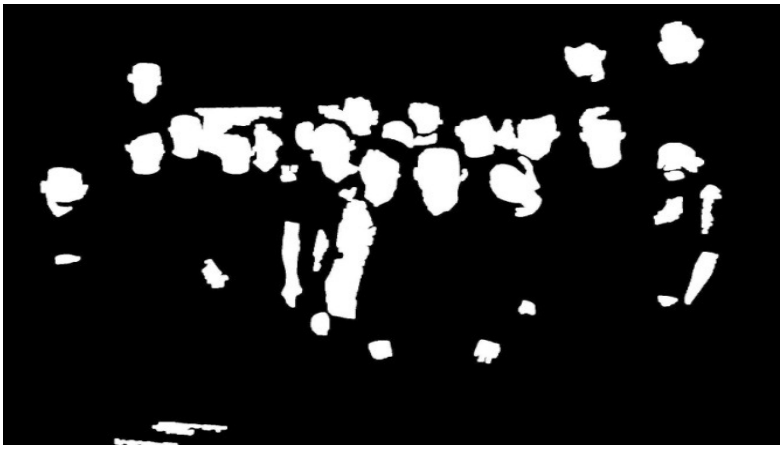


Fig.5. Some non-face regions eliminated binary image

Based on the above regions, bounding box is drawn and the results are shown in Fig. 6. It shall be observed that there are certain non-face regions which is due to the fact that the skin colour is same for both hands and face regions and hence the proposed approach produce certain false face regions. The results are shown in Fig. 6.

Fig. 6. Face and non-face regions with bounding box

Step 4: Removal of non-face regions:

It shall be observed from Fig. 6 that the output image contains many unwanted regions like fist, shoulders, jacket, exposed hands and arms etc. These are removed based on the fact that they occupy lesser area than the other regions or that the aspect ratio does not depict a face region. Thus we have set the following criteria based on which the eliminations of unwanted regions are made:

- Based on area: The area of each region is found out and the threshold is set and all the areas which are above the threshold value are accepted and others are rejected.

- Based on aspect ratio: Some regions have lesser width and greater height or vice versa, which are other than the face.

- Based on region properties: The region properties such as mean, correlation coefficient, kurtosis and moments are computed and clustering is employed to cluster face regions from non-face regions.

Upon employing the above rules, we could be able to eliminate the non-face regions and the final image with face regions identified are shown in Fig. 7.

Fig.7. Image containing only face regions with red colored bounded box.

4. EXPERIMENTAL RESULTS

In this section, we present the experimental results conducted on the dataset which is collected on our own with varied background. We have conducted experimentation using MATLAB tool on Windows platform with 4 GB RAM. Experimental results are conducted on the database which contains nearly four hundred images taken in and around our campus. Some of the sample results obtained due to proposed approach are shown in Table-1.

5. CONCLUSION

In this paper, we have proposed a methodology to detect face regions from complex background based on morphology and skin color information. In the initial stage, segmentation of probable face regions are identified based on skin color information followed by morphological mask processing to identify almost true face regions. Some simple heuristic rules have been introduced to eliminate non-face regions from probable face regions. Experimental results on our database with varied background are conducted to exhibit the performance of the proposed method.

Table-1. Experimental results on images under varied background.

REFERENCES

[1] Ahonen, T., A. Hadid, and M. Pietik¨ainen. Face recognition with local binary patterns. In Proc. of ECCV, 2004.

[2] Dalal, N and B. Triggs. Histogram of oriented gradients for human detection. In Proc. of CVPR, 2005.

[3] Enzweiler, M. and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. IEEE Trans. on PAMI, 31(12):2179–2195, 2009.

[4] Hjelmas, E. and B. K. Low. Face detection: A survey. Computer Vision and Image Understanding, 83:236–274, 2001.

[5] Jin, H., Q. Liu, H. Lu, and X. Tong. Face detection using improved lbp under bayesian framework. In Third Intl. Conf. on Image and Grahics (ICIG), 2004.

[6] Jones, M and P. Viola. Fast multi-view face detection. Technical report, Mitsubishi Electric Research Laboratories, TR2003-96, 2003.

[7] Laptev. Improvements of object detection using boosted histograms. In British Machine Vision Conference, 2006.

[8] Levi, K and Y. Weiss. Learning object detection from a small number of examples: The importance of good features. In Proc. of CVPR, 2004.

[9] Li, Y., S. Gong, and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. In International Conference on Automatic Face and Gesture Recognition, 2000.

[10] Mita, T., T. Kaneko, and O. Hori. Joint Haar-like features for face detection. In Proc. of ICCV, 2005.

[11] Ojala, T., M. Pietik¨ainen, and T. M¨aenp¨a¨a. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on PAMI, 24:971–987, 2002.

[12] Pham, M. -T. and T.-J. Cham. Fast training and selection of haar features during statistics in boosting-based face detection. In Proc. of ICCV, 2007.

[13] Roth, D., M.-H. Yang, and N. Ahuja. A SNoW-based face detector. In Proc. of NIPS, 2000.

[14] Rowley, H.A., S. Baluja, and T. Kanade. Neural network based face detection. In Proc. of CVPR, 1996.

[15] Rowley, H.A., S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. Technical report, School of Coomputer Science, Carnegie Mellow Univ., CMU-CS-97-201, 1997.

[16] Suard, F., A. Rakotomamonjy, A. Bensrhair, and A. Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. In IEEE Intelligent Vehicles Symposium, 2006.

[17] Viola, P and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proc. of CVPR, 2001.

[18] Wu, J., J. M. Rehg, and M. D. Mullin. Learning a rare event detection cascade by direct feature selection. In Proc. of NIPS, volume 16, 2004.

[19] Yan, J., S. Li, S. Zhu, and H. Zhang. Ensemble svm regression based multi-view face detection system. Technical report, Microsoft Research, MSR-TR-2001-09, 2001.

[20] Yan, S., S. Shan, X. Chen, and W. Gao. Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection. In Proc. of CVPR, 2008.

[21] Yang, M, -H., D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. IEEE Trans. on PAMI, 24(1):34–58, 2002.

[22] Zhang, G., X. Huang, S. Z. Li, Y. Wang, and X. Wu. Boosting local binary pattern (LBP)-based face recognition. In Proc. Advances in Biometric Person Authentication, 2004.

[23] Zhang, L., R. Chu, S. Xiang, S. Liao, and S. Z. Li. Face detection based on multi-block LBP representation. 2007.

[24] Zhu, Q., S. Avidan, M.-C. Yeh, and K.-T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In Proc. of CVPR, 2006.

INTENTIONAL BLANK

IRIS BIOMETRIC SYSTEM USING A HYBRID APPROACH

Abhimanyu Sarin

B.Tech EEE (Student), BITS Pilani, Dubai Campus, Dubai, UAE
absarin1004@gmail.com

ABSTRACT

Iris Recognition Systems are ocular- based biometric devices used primarily for security reasons. The complexity and the randomness of the Iris, amongst various other factors, ensure that this biometric system is inarguably an exact and reliable method of identification. The algorithm is responsible for automatic localization and segmentation of boundaries using circular Hough Transform, noise reductions, image enhancement and feature extraction across numerous distinct images present in the database. This paper delves into the various kinds of techniques required to approximate the pupillary and limbic boundaries of the enrolled iris image, captured using a suitable image acquisition device and perform feature extraction on the normalized iris image with the help of Haar Wavelets to encode the input data into a binary string format. These techniques were validated using images from the CASIA database, and various other procedures were also tried and tested.

KEYWORDS

Iris Biometrics, Hough Transform, Daugman's Algorithm, Localization, Haar Wavelets

1. INTRODUCTION

Today, we live in an age where our identity is what defines us more than anything else, but it is also a lot easier to get lost in the midst of the 7 billion people around us. This gives rise to more acute issues- mostly to do with counterfeiting and imitating another's self, a major security predicament .This is where biometric based recognition systems help us in ensuring the safety and security of the things that matter. The identity of a person is demarcated and stored by using an algorithm designed in a way to match the enrolled image(s) when an individual wants to log into the system again.

The most basic concept in any biometric system revolves around the basic processes of acquiring a high-resolution and feature-rich image, followed by detailed analysis of the desired part using image processing techniques and finally matching these details to a given input image. Iris Recognition systems use a very similar methodology.

First developed by Dr. John Daugman in the 1990s[1], who borrowed the idea from Flom and Safir's patented theoretical design, it has been greatly researched on since to make the automated system more efficient and versatile. Some of the main advantages of the this system is the organ itself- the Iris, a doughnut shaped colored structure in the eye is consistent of various features which is inarguably as unique as a fingerprint, if not more - a very rich, random interwoven texture called the "trabecular meshwork"[2]. These features are not chronologically perturbed and

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 149–159, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4914

are genetically incoherent, meaning even twins have different eyes, and are also stable for a lifetime. It is also convenient in a sense that the eye is usually always there with a person, plus since it is an internal muscle, it relaxes when one dies, so even if it is removed it cannot be used fraudulently. Apart from being convenient, clean and secure, it is also very safe by being unobtrusive, as only an image of the eye is taken, not damaging any neural sensors in the eye.

Iris Biometric Identification systems have found major applications around the globe, it is being used in offices as an entry logging system, in passport offices and at airports to associate the visa details of a person upon arrival, and even to enrol the entire population of a country's legal residents and immigrants. The sophistication of the system, along with the very low false rejection rate, has made it reach the pinnacle of biometric security systems by being both reliable and secure.

2. IMAGE ACQUISITION

Before the process of enrolment via image acquisition beings, it is vital to understand how the eye works as a biological organ, as it helps in determining the specifications of the sensor used to capture the image itself.Furthermore, iris recognition is not synonymous with retina scanning, and while both are ocular-based, the former makes use of a high definition photographic detail of the person's iris which is used to examine its unique structure. Retina scanning on the other hand is potentially hazardous as the blood vessels are exposed during examination under relatively high intensity light.

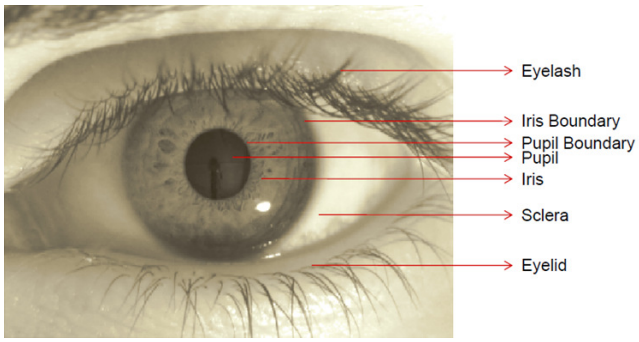


Figure 1. The anatomy of the human Eye, highlighting the Iris and the pupil

The Iris, as discussed in the previous section, is consistent of an extremely feature-rich texture – the trabecular meshwork. The part itself is enveloped by the inner pupillary boundary and the outer limbic boundary around the Iris. This changes in shape with the help of internal muscles in the eye to regulate the contraction and expansion of the pupil under different ambient conditions. The colour of the iris is a result of the melanin content, hence is different shades depending upon its concentration. This colour is not a part of the features to be extracted for examinations, rather the melanin interferes with the image if taken by optical sensors using wavelength of visible light. The melanin content peaks in the UV band in the absorption spectrum at 350nm, and with visible being at 400-700nm [3], the dark eyes (hazel, with high density of melanin) in those bands look despondent because of the low albedo and iris images dominated by the specular corneal reflection (mirror-like) cannot be processed very well. The blue colour pigmentation is naturally a resultant of the long wavelength light penetrating the anterior layer and the stroma. But the far end of the spectrum contains the NIR (Near Infra-Red) region, which has a low absorbance coefficient, hence making pigmentation/melanin density irrelevant. Hence most of the commercial iris recognition camera sensors are NIR, unlike the IR imaging used by retina scanners.

(a) (b) (c)

Figure 2.from left to right (a) photograph of the human eye taken under visible light in constrained/indoor conditions (b) the same eye captured with NIR camera (c) the evidence of corneal reflections, with image taken under visible light in un-constrained environments.

The basic trick in computer vision or imaging systems is to distinguish the difference between Lambertain iris and specular corneal components of reflections. The curved cornea surface obeys the Snell's Law and absorb the ambient light having a broad band of wavelengths. The NIR imaging camera has a blocking filter which cuts off the shorter wavelengths in the remaining band of light allowed to pass through, scrubbing out the ambient reflections. The image acquired by the NIR camera, used for unobtrusive imaging at a distance of say 1metre, is hence much more descriptive of the actual iris image, as even darkly pigmented irises reveal rich and complex features.

3. LITERATURE REVIEW

The history and the significant progress on iris biometric systems dates back to approximately 2001. But the idea itself is over a 100 years old, when the French artists Bertillion[4] mentioned the uniqueness of the features of an iris in his thesis "La couleur de l'iris. Revue scientifique". The automated algorithms were developed with the patent for an unimplemented design by Flom and Safir. They very elaborately designated the use of a highly monitored, occlusion-free environment, using a headrest and a manually operator for a fixed gaze at the subject's eye[5].
The illumination/ ambient properties were changed to determine the change in radius and size of the pupil as it contracted and expanded. They were amongst the first to suggest the various operators and tools in image analysis like corner detection, circular Hough transform and ways to extract the information from the iris. They even mentioned a given threshold for the intensity values to vary within, which was constant for all the images in the database.

The most notable pioneers in Iris Recognition systems with progressive and revolutionary algorithms are- Dr. John Daugman for first patenting a design using the integro-differential operator for Iris boundary localization and the rubber-sheet model for normalization and Wildes et al. developed an algorithm using circular edge maps to compute the boundary and a Hough transform to detect circles.

3.1. Daugman's Algorithm

In 1994, the most stable work on an iris biometric recognition system was evolved from the patent and publications by Dr. John Daugman [1] who described the functionality of this system in acute detail.

The biometric system also evolved with respect to the numerous operators used in the algorithm. Similar to the work done on face recognition systems and the speed cameras seen on the streets,

the eyes were searched using a "deformed template". The primary part of iris localization included the segmentation and clear definition of the pupillary (inner) and the limbic (outer) boundaries. These were defined using a definite operator known as the integro-differential operator, which searched for boundaries in a given parameter space[1], [6].

$$max_{(r,x_0,y_0)} \left| G_\sigma(r) * \frac{\partial}{\partial r} \oint \frac{I(x,y)}{2\pi r} ds \right|$$

Where $G_\sigma(r)$ is thesmoothing function and I(x, y) the image in terms of the representative coefficients of the intensity values of the circular bound region within the x,y parameter space, with x_0, y_0, r being the circular and radial coordinates in the plane. These coordinates are maxed out within the measurement of the pupillary and limbic boundaries defined by the iris and pupil contour, but with the assumption that the potential illumination of the pupil is the maximum gradient circle (practically measured to be 0.8 minimum).

This formula was also experimented with and evolved into a much more sophisticated design, with the inclusion of images with the iris having an off x-axis gaze being a permitted entry[7].
Daugman's algorithm used the rubber sheet model, a method of mapping the external polar coordinates on a circular plane to transform it into a rectangular extracted iris region, irrespective of the noise factors like the eyelids and the eyelashes, which were excluded later.

For feature extraction, he used the 2D Wavelet operators to disintegrate the given image and re-assemble it by marginally reducing the size of the image, without consistently reducing the amount of image stored. After texture analysis, the information is matched using the Hamming distance, which is essentially a difference between the two iris code segments, with the word 'iris code' being coined by Daugman himself as a representation of the iris texture in binary stamp format. He is acclaimed as the father of Iris biometric systems.

3.2. Wilde's Algorithm

Wilde et al. [2]is another prominent scientologist who headed the project at Sarnoff Labs to develop an iris biometric system, with a technical approach slightly distinct from Daugman's. In 1996 and 1998, he had two patents which constituted of a unique acquisition system as well as a slightly less consistent but very effective automated iris segmentation method.

Instead of using a NIR video camera to capture a digitalized image, they have used a standard high resolution camera but with a distinct diffused light source or also described as "a low light level camera".

The iris localization is another distinction. While Daugman used the integro-differential operator, Wildes uses a much more primary route of firstly calculating the binary edge map of the image and then using Hough transform and the relative accumulator function to calculate the intensity levels of pixels constituting a circle, or if the image is distorted by noise, arcs. This algorithm helps detect the pupillary and limbic boundary contours which are then segmented and used the segmented image is sent for feature extraction.

The second distinct method is the usage of a Laplacian of Gaussian filter over multiple overlapping stages in order to produce a template for feature extraction instead of using the compressional methods like Haar wavelet decomposition or 2D Gabor filters.The matching is done by computing the normalized correlation as a measure of the similarity and distinctiveness of two iris codes.

4. IMPLEMENTED METHODOLOGY

The algorithm as a whole is divided into four distinct steps leading up to the matching, using a hybrid approach implemented using the CASIA database V 3.0 Iris Interval and practically verified using MATLAB R2014aStudent version. The system developed consists of the following steps: (1) Image Pre-processing using histogram matching, thresholding and canny edge operator (2) Localization of pupillary and limbic boundary using Circular Hough Transform (3) Iris Normalization using Daugman's rubber sheet model (4) Feature Extraction using Haar wavelets and binary encoding.

4.1. Pre-processing Techniques

Due to the presence of ambient variations in distinct images with varying levels of illumination, the histogram of each individual image needs to be have at least two distinct peaks which represent the distinct greyness intensity level in the image, the darkest part inarguably being the pupil. After selecting a particular image with a bi-modal histogram, the other images are matched taking the former as reference. The matching will usually induce contrast enhancement and the image is blurred later using a Gaussian filter.

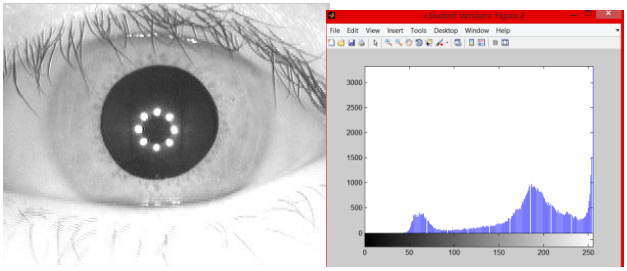
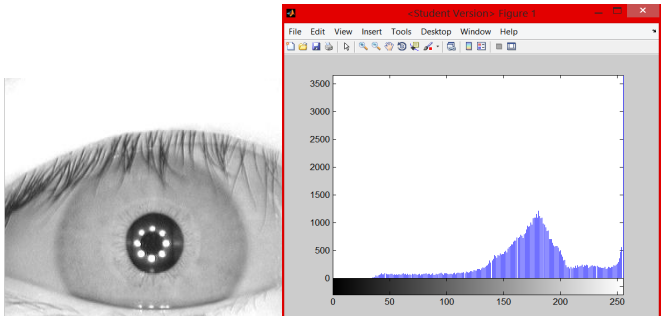
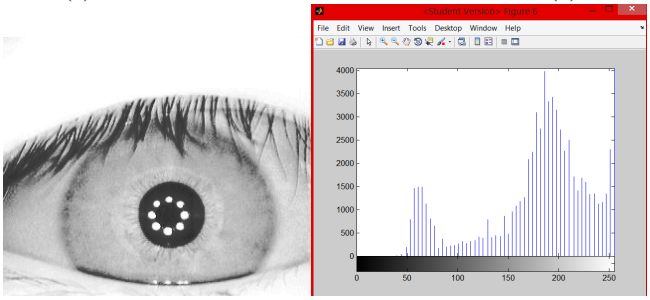


Figure 3. from left to right (a) The reference iris image (b) The histogram of the corresponding unequalized image with distinct peaks



(a) (b)



(c) (d)

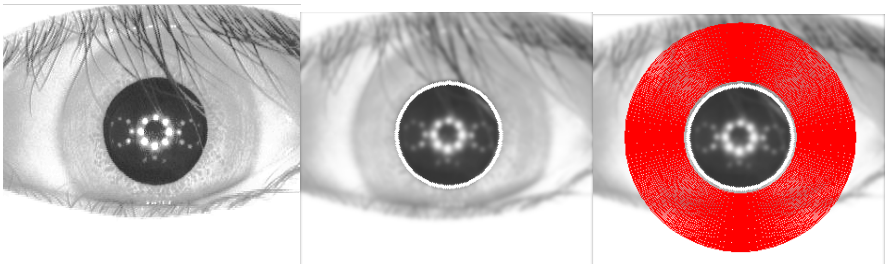
Figure 4. From top left to bottom right (a) The Iris image of the enrolled image (b) the corresponding unaltered histogram (c) The image after smoothing and histogram matching (d) the altered bi-modal histogram

To make sure that the pupillary boundary is properly segmented, a threshold is placed. The threshold is introduced before the canny edge operator is used in order to define only the high transition of intensity levels between the pupillary boundary and the iris [8], [9]. The Gaussian filter helps blur all other noises as well, like the eyelashes and eyelid boundaries.

4.2. Boundary Segmentation

After the image has been thresholded and the histogram equalized, the pupillary boundary becomes easy to localize. The localization is a result of the canny edge operator which establishes the coordinates of the boundary[10]. These coordinates are better defined using circular Hough Transform and the centres and radius of the pupil are noted and stored. Circular Hough Transform (CHT) is transforms a given subset of binary edge points present into an accumulated array of votes in the parameter space. For each edge point, votes are accumulated in an accumulator array for all parameter combinations. The array elements containing the peak number of votes indicate the presence of the shape.

These centric and radial coordinates aid in locating the outer iris (limbic) boundary as well, which is harder to localize as the intensity level of the transition values is not high enough. Since the pupillary and limbic boundaries are almost concentric, the later boundary can be approximated by building concentric circles and the intensity values of the pixels lying over the edge of the boundaries of these circles are calculated and summed using an accumulator array[11]. The difference between the summed values of each consecutive circle is noted and the boundary having maximum variation in intensity as compared the previous one is the limbic periphery. The eyelashes and eyelids, if they are covering the Iris region will not be excluded at this stage.



(a) (b) (c)

Figure 5. from left to right (a) resized iris image (b) image after pre-processing and pupil localization (white) (c) The concentric circles (red) produced by iterative solutions around the pre-detected pupil boundary

The next step is Normalization of the doughnut shaped regionby unwrapping the iris from its polar equivalent's Cartesian coordinates[6], [12].Though the word normalization is often mistaken to be synonymous with equalization or minimization of redundancy in statistical terms, in relations to this subject, normalization refers to the creation of a differently scaled/shifted version of the dataset. Hence, though the data remains the same, it is shifted by functions called pivotal quantities, whose sampling does not depend upon given parameters.The process of localization of an iris image has a major effect on the annular subset from the rest of the image, such that it is not linearized. Daugman had suggested a rubber sheet model, considering the surrounding and acquisition devise to cause the change in pupil's radius by dilation.

The Iris in unwrapped and all points along the edge contour map and within the boundary itself are converted into their polar Cartesian counterparts. The eyelashes and other noises are excluded by using a canny edge operator with a different threshold value.

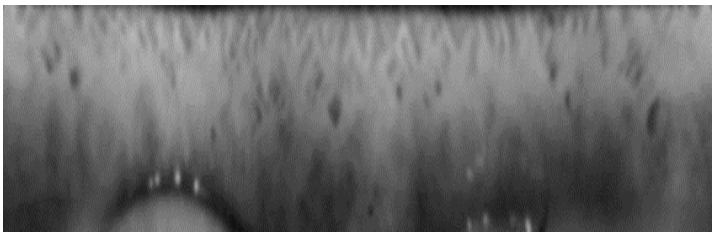


Figure 6. The normalized image obtained after contrast enhancement

4.3. Feature Extraction

The next process done on the normalized iris region is feature extraction [13]using Haar Wavelets[14].

Haar wavelet is a sequence of square-shaped functions which are rescaled as a part of some basis, usually a wavelet family. The sequence itself, very similar to the Fourier series, helps in the basic decomposition of a picture into its constituents, but without the use of sinusoidal functions. They essentially allow us to separate out the low frequencies and the high frequencies via an iterative method to allow for the compression of an image, also called as JPEG compression.

The basic idea behind the compression is to treat the given image (digital) as an array of numbers (matrix).Since the picture say, is a 256x256 pixel grey scale image, the image is stored in the form of a 256x256 matrix, with each matrix element being a whole number ranging from 0 (for black) to 225 (for white). The JPEG compression technique will allow us to keep on decomposing the image from 1x1 to 4x4 to 8x8 and so on till 256x256, assigning a given matrix to each black [15]. There will come an instance that the smaller elements of the decomposed array will have negligible value and hence can be neglected all together, allowing for compression of an image.

In mathematical 1D terms, Haar wavelets have an orthonormal basis for an interval, say, say, L^2[0,1]. When we multiply a unit function with a square function w (x) and integrate, the solution is zero. The compression happens as the time period of the function is squeezed. Eventually, we get smaller functions, removable with minimal loss of resolution [16].

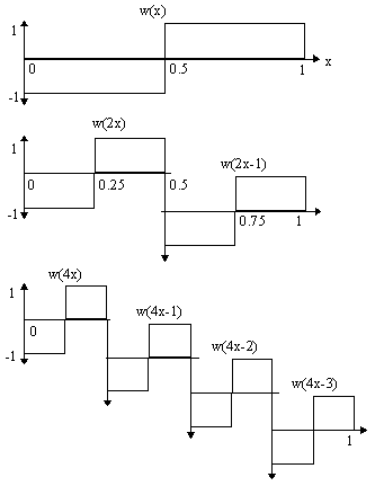


Figure 7. The ‘squeezing’ of an orthonormal function after iterative convolutions

In the figures above, it is clearly demonstrated how a signal, say w(x), which is a square function, when changed, or elongated using w(2x) will give us a similar shape function, but the functions are now squeezed when convolving the same signal to obtain w(2x-1), occupying the same space as the original signal, but compressed as the series progressed over lower frequency regions.
It is apparent that the iterative solutions obtained after constantly decreasing the frequency of the signals will leave us with smaller constituents again, which can be neglected. Though some data is hence lost in the compression technique, it is of very small value and can easily be taken back into consideration when comparing two strings and matching them under a given criteria.

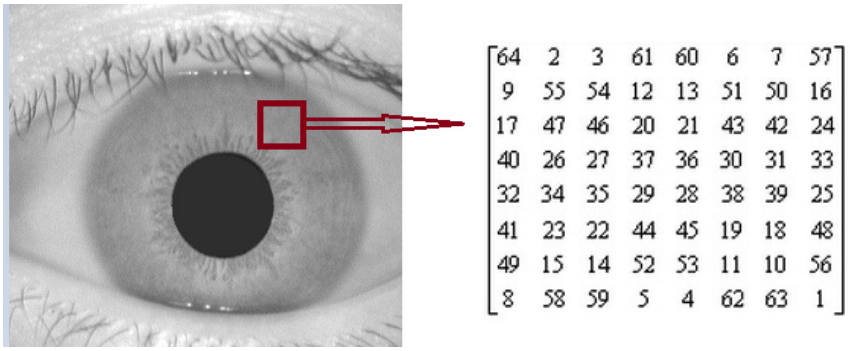


Figure 8. The the constituent elements in form a matrix for a decomposed 2D image

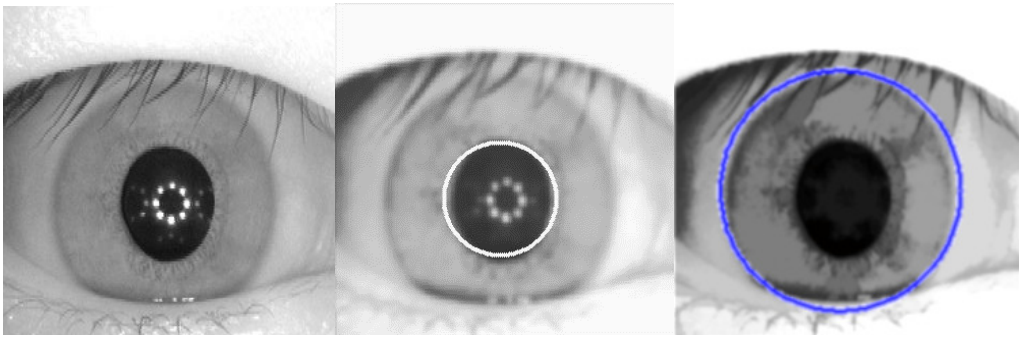
4.4. Binary Encoding

Haar wavelets allow for the compression of the image as well as helps us extract coefficients as approximation, vertical, horizontal and diagonal components. The former coefficients are decomposed by 3 levels and the results obtained as individual arrays from the column and row-wise summationcan be combined to form a singular matrix consisting of various negative and positive values.

To form the feature template, these values can be encoded in the form of binary numbers (1's and 0's), with the positive coefficients, including zero, are taken as 1, while the negative ones are 0. This gives a binary array which can be assigned to each iris image.

5. RESULTS AND DISCUSSION

The Iris biometric recognition systempresented in this paper was tested using the Iris images borrowed from the CASIA database who used an NIR homemade optical sensor to capture images with LED lighting. The procedure was verified using MATLAB R2014a on a 2.6 GHz processor with a dedicated graphics card. The time taken to localize the boundaries and encode the features of the normalized image into binary format was roughly 6.357 seconds per image at an average.



(a) (b) (c)

Figure 9. from left to right (a) An image with an erratic pupillary boundary (b) detection of pupillary boundary using CHT (white) (c) detection of limbic boundary (blue)

The primary results using the above algorithm has a success rate for recognizing and segmenting an iris region is 84% considering a definite but random subset of the database. The data obtained from this iris code was stored in separate folders in the database as the final output. This output is thus enrolled in the system and associated with the subjects' identity. If multiple iris codes from the same or different individuals are present they can be matched and authenticated using B-tree matching, but the algorithm presented in the paper is limited to enrollment and extraction of the Iris.

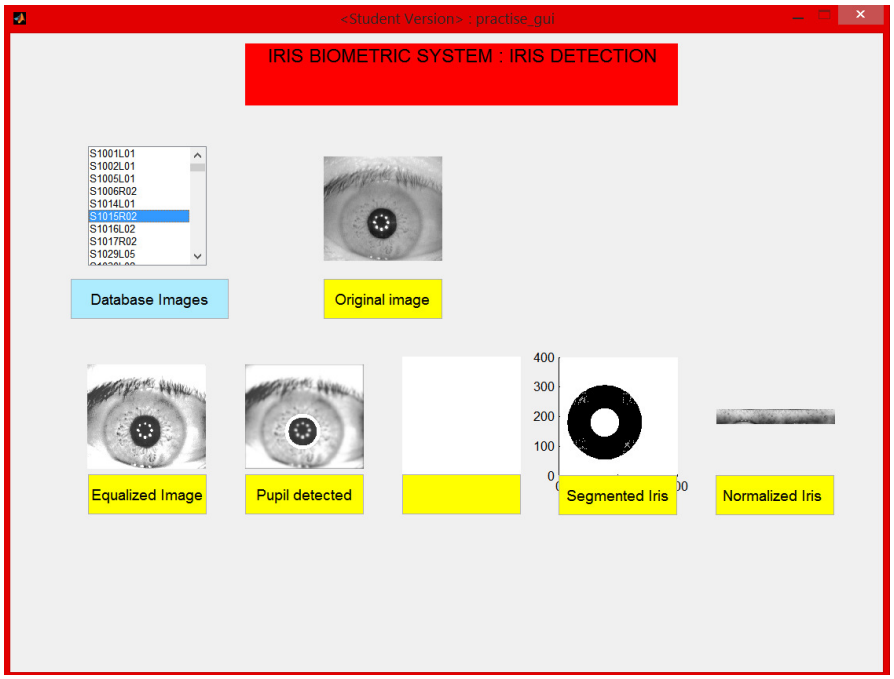


Figure 10. The GUI for the an independently developed Iris Biometric System showing the process till feature extraction

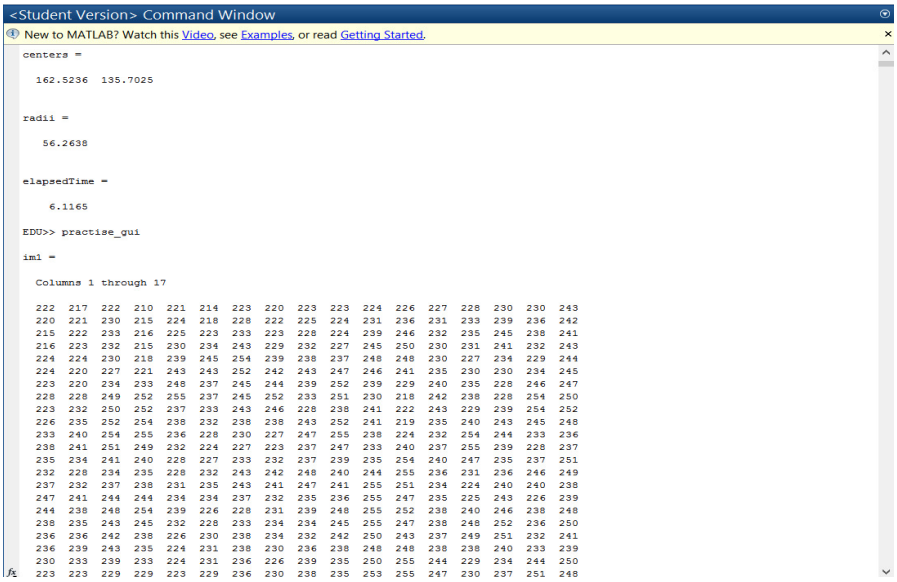


Figure 11. Some results showing the coordinates of boundaries and total elapsed time for an image in the GUI

6. CONCLUSION

The above model and algorithm based system for creating a fairly competent Iris Biometric System was implemented using some basic image processing techniques like histogram equalization and Haar wavelets. The pre-processing methods were particularly important in removing the noise prematurely in order to make sure that they do not interfere with the identification of circular boundaries using Hough Transform. While the process itself just helps us to demarcate the Iris region, we need another method to convert the texture itself and encode it into computer-readable binary format- also called iris codes. The methodology, as proved by the results, is not susceptible to pupil dilation due to varying illumination, specular reflections, or erratic and inconsistent limbic or pupillary boundaries. This system can be validated with the help of B-tree matching with Hamming distance as a matching metric, which allows for comparison of two strings of iris codes. The Hamming distance can be varied with respect to results obtained after computing the rejection rate[17], [18] when comparing two analogous iris codes.

Iris based recognition systems are inarguably a very accurate and precise biometric technique to secure an individual's identity. While the system has itself been praised for being efficient and effective, scientists are still trying to improve the algorithm by publishing new research every year. This approach by using a hybrid method of pre-processing, boundary localization using CHT and feature extraction using Haar Wavelets is also another evolved notion of the same system currently being used all across the globe.

ACKNOWLEDGEMENTS

I would like to express my deep and humble gratitude to Dr. Jagadish Nayak, my guru, who helped me implement this method efficaciously. Also I would like to thank the CASIA team for being kind enough to provide a huge database of Iris Images to the public.

REFERENCES

[1] John Daugman, (1991)"Biometric Personal Identification System Based On Iris Analysis", July, UK
[2] Richard P. Wildes, (1997)"Iris Recognition: An Emerging Biometric Technology", IEEE journal July
[3] John Daugman, (2012) ICB Conference Key Note on Iris Recognition", New Delhi
[4] Bertillon, (1885) "La couleur de l'iris. Revue scientifique", 36(3):65–73.
[5] Hugo Pedro Martins CarricoProenca, (2006)"Towards Non-Cooperative Biometric Iris Recognition", University of Beira Interior, October
[6] John Daugman, (2004)"How Iris Recognition Works" Invited Paper, IEEE Transactions on Circuits and Systems for VideoTechnology, Vol. 14, No. 1
[7] John Daugman, (2007) "New Methods in Iris Recognition", IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, Vol. 37, No. 5
[8] Li Ma, Yunhong Wang, Tieniu Tan, (2002) "Iris Recognition Based on Multichannel Gabor Filtering", ACCV2002: The 5th Asian Conference on Computer Vision, 23--25 January, Melbourne, Australia
[9] PrateekVerma, MaheedharDubey, Praveen Verma, SomakBasu, (2012) "Daughman's Algorithm method For Iris Recognition-A Biometric Approach", International Journal Of Emerging Technology And Advanced Engineering, Issn 2250-2459, Volume 2, Issue 6
[10] Dr. P.K. Biswas, (1999) "Lecture series on Digital Image Processing", Indian Institute of Technology, Kharagpur, NPTEL courses – Joint IIT initiative
[11] NIT Rourkela, "Iris Biometric System", CS635 Dept. of Computer Science & Engineering,
[12] Zhenan Sun, Tieniu Tan, (2009) "Ordinal Measures for Iris Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 12
[13] Presentation slides, "Feature Extraction from Images", LehrstuhlSprachsignalVerarbeitung
[14] Colm Mulcahy, (1995) "Image compression using the Haar wavelet transform", Spelman Science and Math Journal
[15] Kevin W. Bowyer, Karen P. Hollingsworth and Patrick J. Flynn, "A Survey of Iris Biometrics Research: 2008-2010"
[16] Khattab M. Ali Alheeti, "Biometric Iris Recognition Based on HybridTechnique", International Journal on Soft Computing (IJSC) Vol.2, No.4, November 2011
[17] ZaheeraZainalAbidin, MazaniManaf, Abdul SamadShibghatullah, KamaruzamanJusoff, Rabiah Ahmad, ZakiahAyop, SyarulnaziahAnawar, AzizahShaaban and Mariana Yusoff, "A New Hybrid Embedding Method in Iris Biometric System", Australian Journal of Basic and Applied Sciences, 7(3): 46-50, 2013
[18] HunnyMehrotra, "On the Performance Improvement ofIris Biometric System", Thesis report, National Institute ofTechnology Rourkela, Odisha, India, March 2014

AUTHOR

Abhimanyu Sarin was born in New Delhi, India in 1992. He graduated from Birla Institute of Technology and Science, Pilani – Dubai Campus in 2014 with a B.Tech degree in Electrical and Electronics Engineering. His thesis coursework was completed during under graduation on Developing a fast and reliable Iris Recognition System for security purposes.

INTENTIONAL BLANK

THE EFFECT OF APPLYING GAUSSIAN BLUR FILTER ON CAPTCHA'S SECURITY

Ariyan Zarei

Computer Science Student,Shahid Beheshti University, Tehran, Iran
Member of Young Researchers Club, Karaj branch,
Islamic Azad University, Karaj, Iran
arianzarei73@gmail.com

ABSTRACT

Providing security for webservers against unwanted and automated registrations has become a big concern. To prevent these kinds of false registrations many websites use CAPTCHAs. Among all kinds of CAPTCHAs OCR-Based or visual CAPTCHAs are very common. Actually visual CAPTCHA is an image containing a sequence of characters. So far most of visual CAPTCHAs, in order to resist against OCR programs, use some common implementations such as wrapping the characters, random placement and rotations of characters, etc. In this paper we applied Gaussian Blur filter, which is an image transformation, to visual CAPTCHAs to reduce their readability by OCR programs. We concluded that this technique made CAPTCHAs almost unreadable for OCR programs but, their readability by human users still remained high.

KEYWORDS

CAPTCHA, Gaussian Blur, Image Transformations, Optical Character Recognition (OCR), Internet Security

1. INTRODUCTION

Nowadays many people use internet to provide them with their needs such as shopping, banking transactions, registrations, communications, etc. Protecting servers and websites on internet from dangerous threats and attacks has become a serious problem. One of the important and dangerous threats is automated false registrations on webservers that can waste the recourses and finally cause serious damages to the servers.

Many of the registration servers use CAPTCHAs to prevent these attacks. CAPTCHA stands for "Completely Automated Public Turing Test to Tell Computers and Humans Apart" [1]. Actually CAPTCHA is a test that can distinguish human users from robots and programs. Common kinds of CAPTCHAs are audio CAPTCHAs and visual CAPTCHAs.

Most of the websites and registration systems use the visual CAPTCHAs which is usually an image containing sequence of characters with some noises. Websites show this images to users and ask them to enter characters correctly. Because of the weak points of OCR applications this kind of CAPTCHAs has become very common. However, the OCR applications have improved over time and many of them are able to remove the noises and recognize the words.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 161–165, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4915

In this research we have applied Gaussian Blur filter on visual CAPTCHAs and we have investigated that how much they are secured against OCR programs and readable by humans.

2. PREVIOUS KNOWLEDGE

Generally there are various kinds of CAPTCHAs, OCR-Based CAPTCHAs like Gimpy method, Pattern recognition CAPTCHAs like BONGO, and Sound-Based CAPTCHAs.

The BONGO CAPTCHA asks the user to solve a visual pattern recognition problem. It shows 2 series of shapes and patterns with different colors and sizes, then shows another pattern and asks the user to determine that the shown pattern belongs to which of the two series of patterns. [1]

The Sound-Based CAPTCHAs was first designed by Nancy Chan in the University of Hong Kong. It is a system that plays a sound clip containing words and numbers and asks the user to enter what he/she heard. [1]

In this research we worked on the OCR-Based or Visual CAPTCHAs. The idea of visual CAPTCHAs was first created to prevent automated and futile registrations on AltaVista website in 1997. It was done by Andrei Broder and then by DEC Systems Research Center. After that in 2000 Yahoo company decided to have powerful and "easy to use Turing test" to prevent unwanted registrations on its services such as chat room and Email. This system designed by Prof. Manuel Blum at school of computer science at Carnegie Mellon University. [2]

This kind of Turing test first called CAPTHCA by Luis von Ahn, Manuel Blum, Nicholas Hopper, and John Langford from Carnegie Mellon University. [3]

So far visual CAPTCHAs have some common implementations like wrapping the characters via Linear Transformations, Random placement of characters, Noises applied to background, adding horizontal lines over the characters, etc. [4]

The webservers use different CAPTCHAs with different implementations and methods. For example the Gimpy method which is an OCR-Based CPTCHA works by creating an image with colored and noised background containing 7 words with wrapped characters, Then asking the user to enter 3 words out of the 7 words in the picture. [1]

We have other methods but most of them use the common implementations and techniques that we mentioned above.

3. METHODOLOGY

What we have done in this research was applying Gaussian Blur filter to some CAPTCHAs and test their readability by human users and OCR systems.

3.1. Gaussian Blur

Gaussian Blur is a kind of image filter that uses the Gaussian function to make an image blurred. The two dimensional Gaussian function is:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

X is the distance from horizontal axis, y is the distance from vertical axis andσ is the standard deviation of the Gaussian distribution which is also related to Radius of Blurriness in image processing apps.

In order to apply this filter to an image, we have to generate a surface with Gaussian function. The contours of the surface are concentric circles with Gaussian distribution from center point. Values from this distribution build a convolution matrix which should be applied to image to reach final result. [5]

3.2. Applying Gaussian Blur to CAPTCHAs

We have generated 50 simple OCR-Based CAPTCHAs with simple white background and simple characters. These CAPTCHAs contained two meaningless 4-7 letters word separated by space. We created these images by an application that we've written in C# language. In this application we generate two random word as mentioned above and then we draw it to an image.

Then we applied Gaussian filter on these images with Radius of 1 and Radius of 2 by using another application in C# as you can see it in figure2. In this application we surveyed on the image's pixels and applied the convolution matrix to each pixel. You can see some examples of these CAPTCHAs in figure1.

Figure1. Visual CAPTCHAs with Gaussian Blur filter applied on.
Left: Gaussian Blur with radius of 1
Right: Gaussian Blur with radius of 2

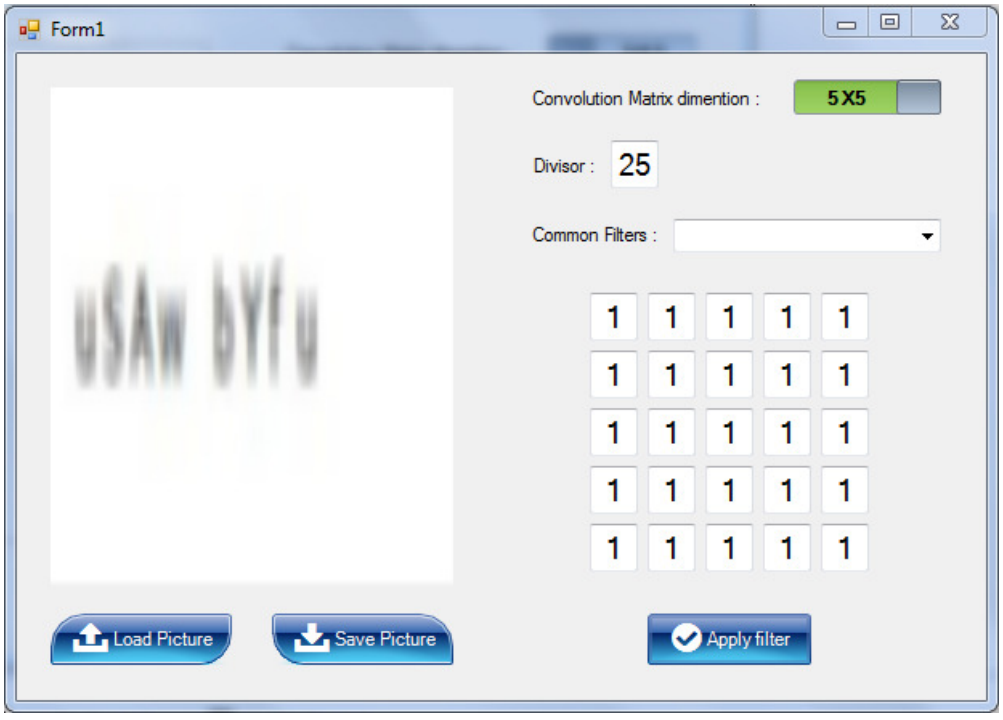


Figure2.the application that was created to apply Gaussian filter to the images

To apply the Gaussian filter to an image in PHP (as a common web server language) there are two ways, we can use “image filter” function [6] or we can implement the algorithm that has been explained above manually. To do so, we can use the “image colorat” function [7] to read each pixel’s color and “image setpixel” function [8] to set the color of that pixel which obtained from applying the convolution matrix to it. The rest of the implementation is just loops and simple mathematical operations.

At first we have tested the readability of these two series of CAPTCHAs on human users by showing them the CAPTCHAs and asking them to enter what they saw. Also we asked them to rate the readability of these CAPTCHAs by giving a number between 1 and 10. We reached to satisfying results in this test. They were able to accurately distinguish 92% of the blurred CAPTCHAs with radius of 1 and 90% of them with radius of 2 and for readability rating we reached the average of 9.2 out of 10.

After that, we tested the security of CAPTCHAs against 2 OCR programs. These OCR programs were ReadIRIS 14 and Free OCR. In some cases they were unable to recognize text from background especially for the CAPTCHAs with radius of 2. In other cases mostly they were unable to recognize the exact words. You can see the complete results in Table 1 below.

Table 1. Results

Humans Results	
Average characters similarity (radius of 1)	99.11%
Average characters similarity (radius of 2)	99.01%
Average exact match (radius of 1)	92.00%
Average exact match (radius of 2)	90.00%
Average readability rating provided by human testers	9.2
OCR results	
ReadIRIS Application	
Average characters similarity (radius of 1)	31.86%
Average characters similarity (radius of 2)	3.45%
Percentage of readable CAPTCHAs (radius of 1)	56.00%
Percentage of readable CAPTCHAs (radius of 2)	30.00%
Average exact match (radius of 1)	16.00%
Average exact match (radius of 2)	0.00%
Free OCR Application	
Average characters similarity (radius of 1)	4.68%
Average characters similarity (radius of 2)	0.52%
Percentage of readable CAPTCHAs (radius of 1)	100.00%
Percentage of readable CAPTCHAs (radius of 2)	32.00%
Average exact match (radius of 1)	0.00%
Average exact match (radius of 2)	0.00%
Total Results	
Total average characters similarity on OCR programs	10.13%
Total average exact match on OCR programs	4.00%
Total average characters similarity on humans	99.06%
Total average exact match on humans	91.00%
Radius of 2 results	
average exact match on OCR programs	0.00%
average exact match on humans	90.00%

4. CONCLUSION

Because of some weaknesses of OCR programs, many webservers use OCR-Based CAPTCHAs to prevent futile registrations. In this paper we investigated the effect of using Gaussian filter on CAPTCHAs security. Actually we applied Gaussian blur filter on CAPTCHAs to improve their safety against OCR programs.

Based on the result we acquired (Table 1) the CAPTCHAs with Gaussian Blur filter applied on, are very powerful against OCR programs and also their readability by human users are extremely high. We generated two series of Blur CAPTCHAs, one with radius of one and the other with radius of two. Considering the results, blur CAPTHCAs with radius of two are more efficient than the other one. OCR programs couldn't recognize any of the CAPTCHAs with radius of two however human could recognize 90% of them. So they can be used in webservers to prevent abuse and unwanted registrations on them.

ACKNOWLEDGMENT

We offer our thanks to the president and faculty members of Computer Science Department of Shahid Beheshti University.

Especially thanks to Young Researchers & Elite Club that provided the research possibility for me.

REFERENCES

[1] L.Von et al.,"Telling Humans and Computers Apart Automatically," COMMUNICATIONS OF THE ACM, Vol. 47,No. 2,Feb. 2004.
[2] H.S. Baird and K. Popat, "Human Interactive Proofs and Document Image Analysis," in 5th IAPR International Workshop on Document Analysis Systems, Princeton, NJ, 2002, pp. 507-518.
[3] M. H.Shirali-Shahreza and M.Shirali-Shahreza "Persian/Arabic Baffle text CAPTCHA,"J.UCS, vol. 12, no. 12,pp. 1783-1796, 2006.
[4] A.Hindle, M.W.Godfrey, R.C. Holt "Reverse Engineering CAPTCHAs," in 15th Working Conference on Reverse Engineering, Antwerp, 2008.
[5] Wikipedia The free encyclopedia, http://en.wikipedia.org/wiki/Gaussian_blur
[6] PHP website, http://php.net/manual/en/function.imagefilter.php
[7] PHP website, http://php.net/manual/en/function.imagecolorat.php
[8] PHP website, http://php.net/manual/en/function.imagesetpixel.php

INTENTIONAL BLANK

A 5.99 GHZ INDUCTOR-LESS CURRENT CONTROLLED OSCILLATOR FOR HIGH SPEED COMMUNICATIONS

Chakaravarty D Rajagopal[1], Prof Dr.Othman Sidek[2]

[1,2]University Of Science Malaysia, 14300 NibongTebal, Penang. Malaysia
[1]drchakra@pcocd2.intel.com
[2]othman.cedec@usm.my

ABSTRACT

This paper presents the design of five-stage current controlled inductor-less ring oscillator that were simulated in Silterra 0.18um CMOS Technology with oscillation frequencies up to 5.99 GHz. The design uses cross coupled MOS devices along with active inductor (thus inductor-less) and controlled by current source to aid in switching speed and to improve the noise parameters. The simulations show that the five-stage oscillator achieves frequency in the range of 3.78GHz to 5.99GHz. The simulated phase noise of the same design was -115.67 dBc/Hz at 1MHz offset with a center frequency of 5.99GHz.

KEYWORDS

VLSI &CMOS, LC Oscillators, Phase Noise, Current Source, Active Inductor

1. INTRODUCTION

The Phase-locked Loop (PLL) is a critical component in many high-speed systems since it provides the timing basis for functions such as clock control, data recovery, and synchronization. The voltage/current controlled oscillators (VCO/CCO) is perhaps the most crucial element of the PLL because it directly provides output clock of the PLL.

Any CMOS oscillators can be built using ring structures, relaxation circuits, or an LC resonant circuit. The LC design has the best noise and frequency performance due to the large Q factor of the resonant networks [1]. However, LC circuit in CMOS process increases the cost and the complexity of the chip and also often time creates problems in controlling the eddy current.
On the other hand, oscillators with ring structure are easily built on any CMOS process and it is less complex and costly. The design is also very straight forward and it is also capable of providing multiphase outputs and a wide tuning range. Fig. 1 shows the conventional five-stage ring oscillator. The downside of this ring oscillator is compromised noise performance due to the missing passive LC network. In this article, we present a design that improves the overall characteristics of CMOS ring oscillators to be comparable to those of LC designs by replacing the passive LC network with the active version. The design also adds a current source instead of voltage source to increase the switching speed.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 167–172, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4916

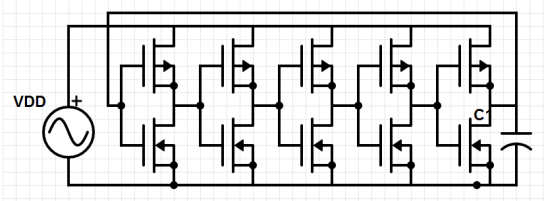


Fig. 1Conventional Five-Stage Ring Oscillator

2. CROSS COUPLED DELAY CELL WITH ACTIVE INDUCTOR LOAD

Fig.2(A) below shows the typical cross coupled delay cell with passive inductor load. These passive inductors in this circuit are realized using an on-chip spiral layout which suffers from huge area consumption, small inductance and strong interaction with the substrate. There are many ways to synthesize an inductor. Self-biased active inductors are one of them. There were some initial researches proposed for MESFET [2] and later re-developed for CMOS [3, 4]. Fig. 2(B) shows one such proposal.

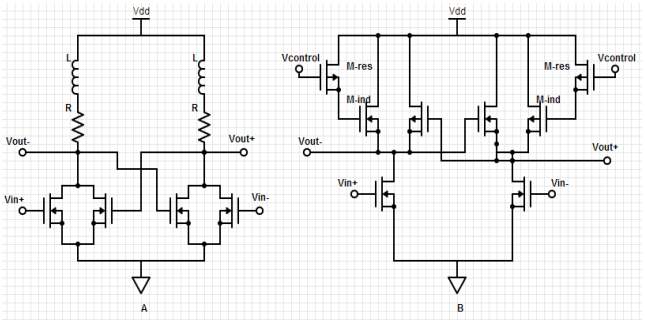


Fig. 2 VCO delay cell. (A) delay cell with passive inductor. (B) delay cell with active inductor.

M-res and M-ind forms the active LRC network. The circuits in Fig. 2(A-B) are simulated with square wave input and the output response isshown in Fig. 3. Note that the output of the delay cell with active inductor (B) is very similar to the one with the passive inductor (A).This is a huge achievement in replacing the passive inductor with the active inductor.

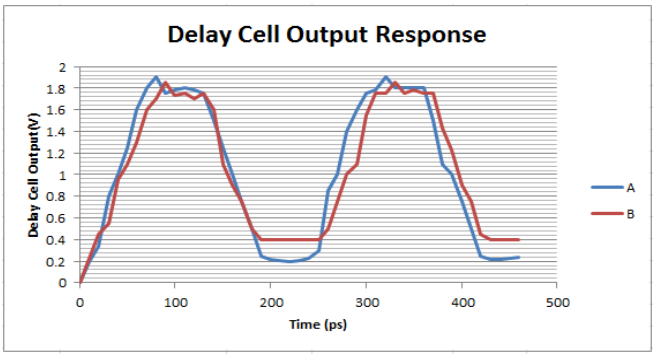


Fig. 3 Output voltage of VCO delay cell.
(A) Delaycell with passive inductor.
(B) Delay cell with activeinductor.

3. CURRENT MODE TECHNIQUE TO IMPROVE PHASE NOISE

The oscillation frequency of a CMOS inverter ring oscillator can be tuned either by adjusting its core supply voltage V_{DD} or its core supply current I_{DD} as shown in Fig. 4 below.

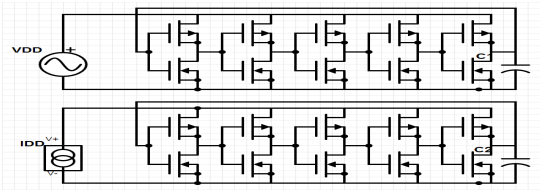


Fig. 4. Conventional Ring Oscillator with voltage modesupply (top)
and current mode supply (bottom).

Phase noise response for the five-stage conventional ring oscillator supplied with V_{DD} and I_{DD} are shown in the Table 1 below.

Table 1. Simulation results of Phase Noise at 1MHz offset
in Voltage Mode and Current Mode.

V_{DD} (V)	2.3	1.8	1.3
Phase Noise (dBc/Hz)	-88.7	-81.8	-79.1
I_{DD} (mA)	5.4	3.5	1.6
Phase Noise (dBc/Hz)	-92.3	-85.4	-82.7

It is clearly noted that the phase noise of the current controlled five-stage oscillator improves by about 4dBc/Hz as compared to voltage controlled. Silterra 0.18um CMOS technology was used in the design and simulation of the ring oscillator. To realize the current mode, PMOS current mirror technique implemented by using 3.3V I/O transistors. Fig. 5 depicts this.

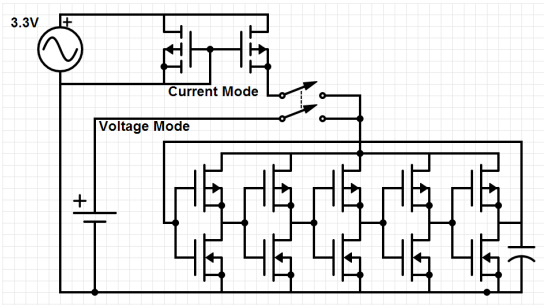


Fig. 5 Conventional ring oscillator implementing PMOS current mirror
and the connections for voltage mode and current mode supply.

4. PROPOSED INDUCTOR-LESS CURRENT CONTROLLED OSCILLATOR (ICCO)

Using the techniques described in Section 2 and Section 3 along with the negative delay skew techniques [5,6], a novel oscillator design has been proposed as shown in Fig. 6. Each stage of the

oscillator comprises of dual-delay cell with active inductor load scheme as depicted by Fig.2(B) and the PMOS current mirror for the current mode.

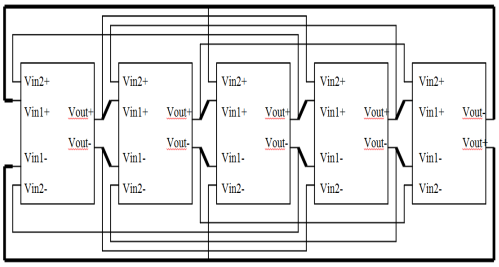


Fig. 6 Inductor-less Current Controlled Oscillator (ICCO)

The oscillator is laid out in Silterra 0.18um CMOS technology and later the parameters were extracted for simulation purposes. Fig. 7 shows the layout of the oscillator.

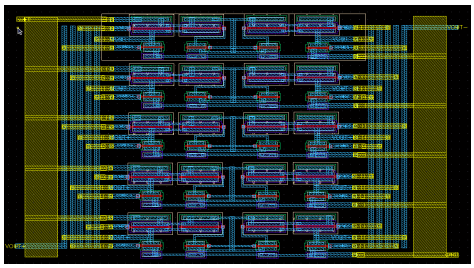


Fig. 7 Layout of Five-Stage ICCO.

The extracted device parameters were simulated using the same 0.18um process. The simulation reveals a peak oscillation frequency 5.99GHz and simulated phase noise of -115.67dBc/Hz at the oscillation frequency of 5.99GHz at 1MHz offset as shown below in Fig. 8. The driving capacitance for this oscillator has been set to 0.5pF for each stage and the supply for the current mirror has been set to 3.3V.

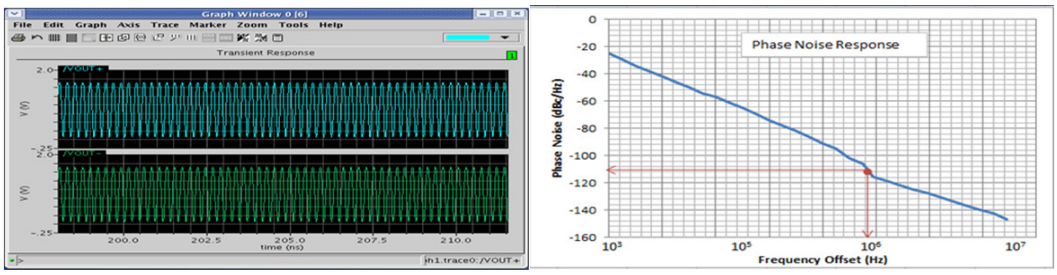


Fig. 8Simulated ICCO Oscillation of 5.99GHz and phase noise of -115.67dBc/Hz at 1 MHz offset.

Numerous simulations were carried out with various I_{DD} settings. Table 2(A-B) below shows the simulation setup and the simulated output frequency and the phase noise of the ICCO for various setting of the I_{DD}.It is quite apparent that I_{DD} needs to be above 2mA for decent phase noise performance. Thus only a current mode circuit can provide such a large current for comparable voltage mode MOS device configurations.

Table 2. (A) Simulation Setup . (B) Simulated Oscillation Frequency and Phase Noise (at 1MHz offset) of the ICCO for various I_{DD}

Simulation Setup	
Process	0.18um CMOS
Voltage	1.8v
Current Mirror Voltage	3.3v
Driving Capacitance	5pF per stage
Simulation Tool	Cadence's Spectre

(A)

I_{DD} (mA)	5.5	4.5	3.5	2.5	1.8
Osc Frequency (GHz)	5.99	5.45	5.16	4.97	3.78
Phase Noise (dBc/Hz)	-115.67	-113.45	-111.2	-107.89	-87.25

(B)

Table 3 below shows the comparison of this work against the other well-known works. It is clearly noted that this work excel in the oscillation frequency along with comparable phase noise. It is very interesting to note that the overall CMOS characteristics and process has improved dramatically when compared to the reported works especially [11] since it achieves almost same output frequency as this work. However, this work is much smaller in layout size (hence cheaper) since it is using active inductor as compared to passive inductor by [11] thus proving characteristics of CMOS has been improved through the replacement of passive inductor with active inductor.

Table 3. Comparison of this work against others

Ref	Technology (um)	Vdd (V)	Phase Noise (dBc/Hz)	Freq (GHz)
[7]	0.18 CMOS	1.8	-118.00	3.0
[8]	0.18 CMOS	1.8	-110.00	1.6
[9]	0.18 CMOS	1.8	-122.90	1.6
[10]	0.18 CMOS	1.8	-109.40	1.5
[11]	0.18 CMOS	1.8	-101.67	6.01
This Work	**0.18 CMOS**	**1.8**	**-115.67**	**5.99**

5. CONCLUSIONS

This paper proposes a current controlled oscillator for an improved frequency oscillation and with active inductor load for an improved phase noise. Achieving a phase noise of -115.67 dBc/Hz at 1MHz frequency offset and a peak oscillation at 5.99GHz is quite obvious that this is capable of being used in high speed communications applications.

REFERENCES

[1] B. Razavi, "A study of phase noise in CMOS oscillators", IEEE J. Solid-State Circuits, vol. 31, pp.331 -343 1996

[2] S. Hara, T. Tokumitsu, T. Tanaka and M. Aikawa, "Broadband monolithic microwave active inductor and its application to miniaturized wide-band amplifiers." IEEE Trans. Microwave Theory and Appl., vol. 36, no. 12, pp. 1920-1924, 1988.

[3] E. Sackinger and W. Fischer, "A 3-GHz 32-dB CMOS limiting amplifier for SONET OC-48 receivers." IEEE J. Solid-State Circuits, vol. 35, no. 12, pp. 1884-1888, 2000.

[4] S. Song, S.Park and H.Yoo, "A4-Gb/s CMOS clock and data recovery circuit using 1/8-rate clock technique." IEEE J. Solid-State Circuits, vol. 38, no. 7, pp. 1213-1219, 2003.

[5] S.-J. Lee, B. Kim, and K. Lee, "A novel high-speed ring oscillator for multiphase clock generation using negative skewed-delay scheme", IEEE J. Solid-State Circuits, vol. 32, pp.289 -291 1997

[6] C. H. Park and B. Kim, "A low-noise, 900-MHz VCO in 0.6-µm CMOS", IEEE J. Solid-State Circuits, vol. 34, pp.586 -591 1999

[7] L. Lu, H. Hsieh, Y. Liao, "A wide tuning-range CMOS VCO with a differential tunable active inductor", IEEE Transactions on Microwave Theory and Techniques, Vol. 54, No. 9, pp. 3462–3468, 2006

[8] A. Tang, F. Yuan, E. Law, "A new CMOS active transformer QPSK modulator with optimal bandwidth control", IEEE Transactions on Circuits and Systems II: Express Briefs, Vol. 55, No. 1, pp. 11–15, 2008

[9] A. Tang, F. Yuan, E. Law, "Class AB CMOS active transformers Voltage-Controlled Oscillators", ISSSE '07, International Symposium on Signals, Systems and Electronics, pp. 501–504, Montreal, 2007

[10] A. Tang, F. Yuan, E. Law, "Low-noise CMOS active transformer voltage-controlled oscillators", MWSCAS 50th Midwest Symposium on Circuits and Systems, pp. 1441–1444, Montreal, 2007

[11] N. Yangqing, C. Baoyong, W. Zhihua, "A CMOS LC VCO with 3.2-6.1GHz tuning range," Chinese Journal of Semiconductors, vol. 28, no. 4, pp. 526-529, 2007.

AUTHORS

Chakaravarty D Rajagopalis currently a Member of Staff Engineer at Intel Corporation USA. He has been with Intel since 1995. He earned his Masters of Science in Microelectronics from University Science Malaysia in 2003 and currently working on his PhD. His current research interest includes VLSI, Microelectronics, CAD tools and various SoC verifications such as Gate level Simulations, Clock Domain Crossings, Synthesis, Formal Equivalence Verification and Static Timing Analysis.

Prof. Dr. Othman Sidek is currently a Prof at University Science Malaysia whose journey as an academian began 20+ years back. He is an esteemed Malaysian scientistand his charisma, talent and skills have enabled him to make meaningful contributions to Malaysia's nation building efforts. He earned his PhD in Information Systems Engineering from Bradford University, UK after completing Masters in Communication Engineering from UMIST, UK. His current research areas mainly focus on Micro-Electro Mechanical Systems (MEMS), Wireless Sensor Network, Embedded System/SoC and VLSI/ASIC Designs.

MULTI-USER SERVICE PLATFORM DESIGN FOR SMART TV & N-SCREEN SERVICES IN OPEN CLOUD ENVIRONMENT

JuByoung Oh[1] and Ohseok Kwon[2]

[1]Koino,Inc, Seoul, Korea (South),
jboh@koino.net
[2]Computer Science Engineering Department,
Chungnam National University, Daejeon, Korea (South)
oskwon@cnu.ac.kr

ABSTRACT

Smart TV has been discussed as a promising device of Post PC category to handle various user needs by adding computing power to general TV. Smart TV is already commercialized and used in web-surfing, on-demand requests on movies combined with Internet enabled set-top box device. There has been specific approach to increase its usability by adding TV apps for specific Smart TV hardware. However, as Post PC perspective, current Smart TV system and architecture are lack of flexibility and need new paradigm. The architecture should provide office-work friendly environment, cover various OS-dependent users and apps based on Android OS & iOS together, and support legacy IT resources. Thus, we propose new platform design to achieve the goal to make Smart TV as a Post PC device based emerging cloud virtualization and N-Screen technologies.

KEYWORDS

Cloud Virtualization, Smart TV platform, Post PC, N-Screen

1. INTRODUCTION

Smart TV is an improved form of legacy TV and has been discussed as one of promising devices for Post PC. Up to now, Smart TV is gradually changing its system architecture by adding functions to increase its usage and coverage. However, previous approaches were insufficient because they were lying on the legacy broadcasting paradigm or dependent on hardware. We suggest new platform design to add more flexibility and to cover weak points of the previous systems.

1.1. Legacy Smart TV & its limitation

The original concept of Smart TV was started to add functions like Internet and Web2.0 specification to legacy TV and it was believed that it would take the role of PC. [1][2][3] Based on the fundamental Smart TV concept, legacy Smart TV system architecture consists of the server providing contents and applications, set-top box clients for home appliances, and reasonable network devices with Internet connection. Even though it had been improved its system and functions continuously, the independent Smart TV system was requested to upgrade its overall

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 173–185, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4917

system architecture because of lack of applications, device-oriented set-top-box, inflexible UI inconvenience, etc.

The following Figure 1 shows the legacy architecture of Smart TV system. It consisted of basic network / broadcasting function controlling engine, UI & overall management module for user interface, codec modules for videos, and web-browsing module to read simple documents and pictures (might be limited). The system can process contents of only video and image which are already pre-defined or set as a standard. Legacy Smart TV platform was usually designed on a closed private environment and needed customization for each company. It was hard to add functions and difficult to change its structure. To cover the weakness of the legacy system, several approaches were being introduced. [4][5][6]

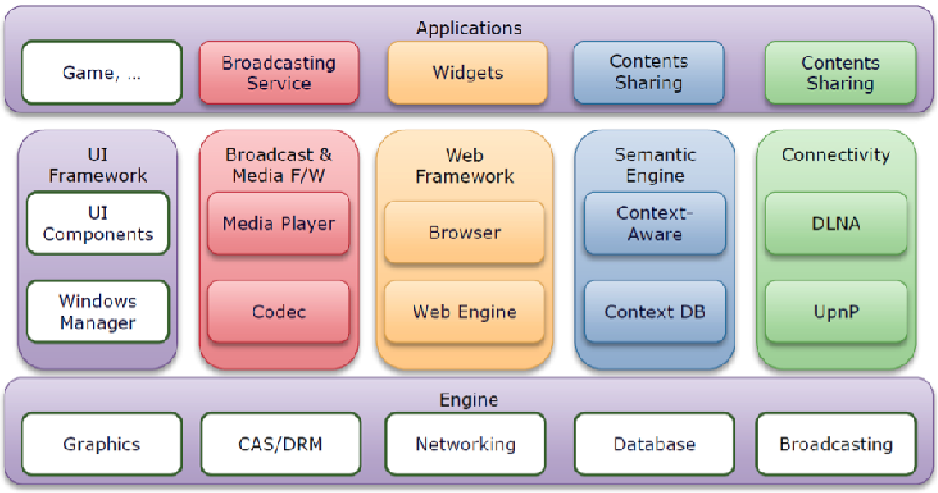


Figure 1. Legacy architecture of Smart TV

1.2 N-Screen & apps trend reflected on Smart TV

Recently new types of Smart TV approaches were introduced by renowned IT companies like Apple, Google, and Samsung to overcome weakness and restriction of legacy Smart TV system.[7][8][9] Those were iTV of Apple, Android TV 2.0 of Google, and Smart HUB of Samsung. According to the advent of these brand new system architecture and infrastructure with cloud computing environment, they anticipated that Smart TV would be a core element of killer contents & applications in IT resources with to the gradual increase of smart devices.

Android OS and iOS smart devices are very common personal devices and also have steadily growing numbers of apps of covering various genres and versatile subjects. Regarding mobile apps, iOS apps were exceeded 700 thousand in 2013 and Android apps were also exceeded 700 thousand at the same time. Each new Smart TV system targeted to be a rich-content Smart TV and to give a strong impact to the industry by providing a lot of apps for customers to feel much more added values compared to that of the legacy Smart TV system of having simple broadcasting capability.

With use-ready apps, contents-transferring cloud platform, and its own brand set-top box, it was tremendous paradigm change of providing rich customer experience and additional side effects compared to legacy Smart TV. Two big software companies' approaches were very similar in that they utilized their own apps and their own smart devices. Unlike two big software companies, Samsung's Smart TV approach was rather TV device oriented approach. Samsung gave

additional value to only the buyers of their Smart TVs by providing their apps only working on theirs. Samsung's approach was not a fundamental change but to give a value to its hardware. They have about 1,500 apps for the TV as of year 2013. Samsung's approach is a trend but it is not the main trend at this moment.

Figure 2. Current Smart TV approaches: Android TV, Apple TV, and Samsung TV

1.3 Cloud computing reflected on Smart TV

Cloud computing is an architecture to provide IT functions as service as like people can use ATM easily even though they have no knowledge on internal solution and used technology. The definition of IEEE is abstracted in one phrase: a paradigm of data stored permanently in a server residing on the network and temporarily in client devices like desktop, tablet, wall-mountable computer, and portable device. Figure 3 shows the concept of cloud computing.

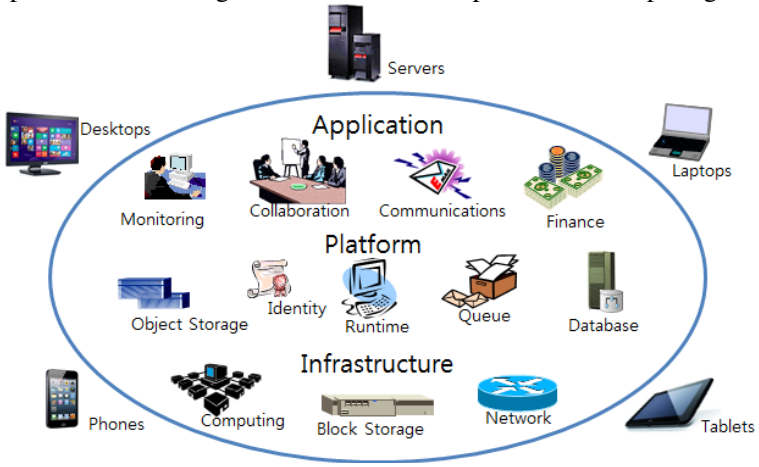


Figure 3. The concept of cloud computing

Comparing to the architecture of independent server and desktop, cloud computing service can reduce the initial purchasing cost and provide mobility to users. It is also a good approach to Green IT by increasing the effectiveness. As a client perspective in a concept of N-Screen, it is

possible to use multiple devices on the same contents without such limitation of OS and location even though there is cross-platform issue. It is also much safer to keep all the user data to the server, not to his own carry devices.

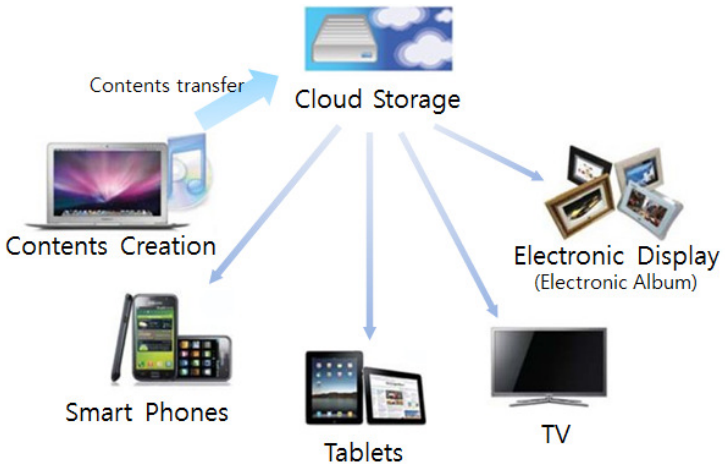


Figure 4. N-Screen concept diagram

By applying the cloud computing technology to Smart TV system, it surely is able to increase the effectiveness of legacy server-client Smart TV structure. Recently, cloud computing integrated N-Screen infrastructure is getting more popular according to the spread of smart devices and better ubiquitous network environment. N-Screen is a good starting point of consideration to apply it to Smart TV platform for user experience enhancement.

As a back-end server side of the Smart TV platform should be definitely cloud based computing. Because Smart TV platform has a lot of apps and contents (videos, etc.), the server system must have capability of effective management. The cloud technology could enhance management effectiveness by sharing resources. The cloud technology also could easily provide flexibility and scalability to the platform. For actual example, Apple has its own cloud space named iCloud and Google has the same kind of cloud space named gCloud.

1.4 Weakness of iTV and Android TV

- *Hardware and OS dependency, Limitation of document processing and individual OS*

Though they are very good platforms to apply for Smart TV, each has both weakness and shortcoming. Each solution is based on its own ecosystem and its own specific hardware devices. iTV set-top box is packed iOS device and is able to share content with iPhone, iPad with iCloud, and to communicate with iStore environment. It cannot use Android OS based smart devices and Google's cloud environment. Though iTV can increase easily user experiences with their apps, it is limited to their own apps and cannot use Android's apps. It is the same situation in Android TV. They cannot use iOS smart devices and iStore's apps, too. As a customer perspective, if he has an Android tablet and wants to see iTV, he should buy forcefully iPad.

Both iTV and Android TV have another weak point if a customer wants to extend it to desktop environment. The two solutions are not interoperable with Microsoft applications on desktop PC working environment. Even though there are some alternative apps to cover it, they cannot cover the most widely used desktop document applications like Excel, Word, and PowerPoint.

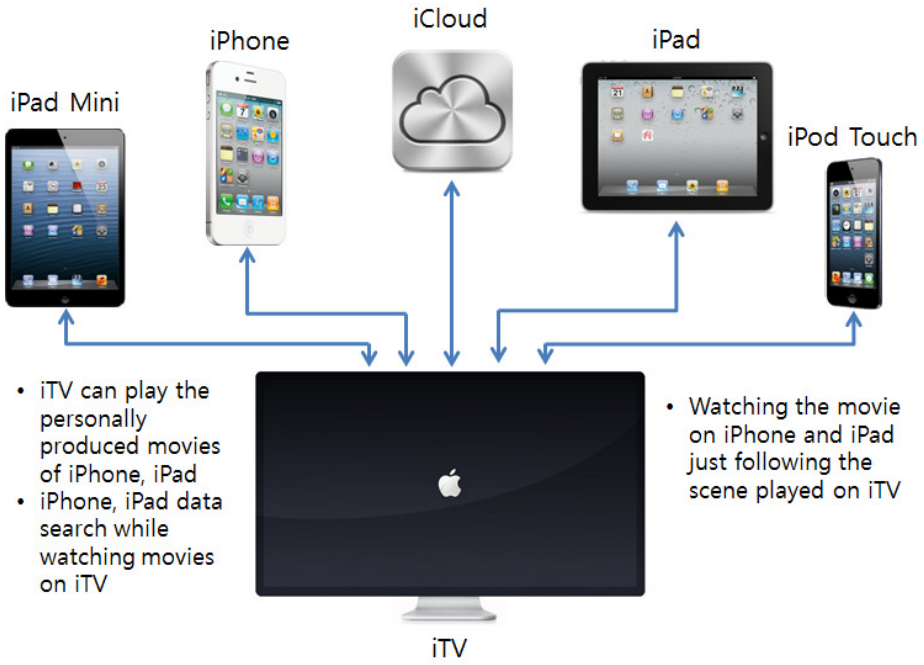


Figure 5. Apple's Smart TV – iTV

Furthermore, Android OS and iOS were originally designed for each individual and were not targeting multi-user or multi-processing. It might be a severe limitation when the OS has to work on simultaneous tasks in Smart TV. When we consider cloud desktop virtualization, it might be also reluctant to adopt mobile OS as one of virtualization guests in that lack of resource management and multi-processing. To increase effectiveness, we must consider deploying multi-user OS as a guest. Well-known multi-user operating systems are Microsoft Windows, Linux, and UNIX. Considering the performance, environment, and cost for it, it should be an effective approach to deploy Windows OS or Linux on x86 hardware.

2. CONSIDERATIONS ON FLEXIBLE SMART TV ARCHITECTURE

Upon the introduction of Android TV and iTV owing to the combination of cloud and smart device technology, Smart TV's capability and user experiences are improved much further comparing that of traditional legacy Smart TV system. However, the approaches have weakness of OS & hardware dependency, lack of document work functionality, and limitation of cross-platform cooperative work functionality. To overcome weaknesses and shortcomings, we considered to deploy the cloud virtualization technology for getting rid of OS barriers, to design multi-purpose VDI (Virtual Desktop Infrastructure) protocol for effectiveness, and N-Screen technology for user accessibility.

2.1 Considerations on Cloud Virtualization technology

Cloud computing itself is already a common terminology to people and is usually considered as a representative example of paradigm change. Recently there are a few approaches to deploy VDI (Virtual Desktop Infrastructure) with cloud computing platform. [10][11][12][13] VDI concept is to integrate desktop computers into the cloud platform. When they apply VDI to cloud computing system, they are able to not only use resource effectively but also utilize zero client or thin client as a client device. It also gives big advantages to operate and manage desktops and to keep

desktop data in secure. There are several trials to upgrade effectiveness of using virtualization resources: CPU, memory, HDD space, and processes. To apply VDI to cloud computing environment, there needs hypervisor which can control virtualization guest OS. Most well-known approaches in OSS (Open Source Software) are OpenXen and KVM. [14][15]

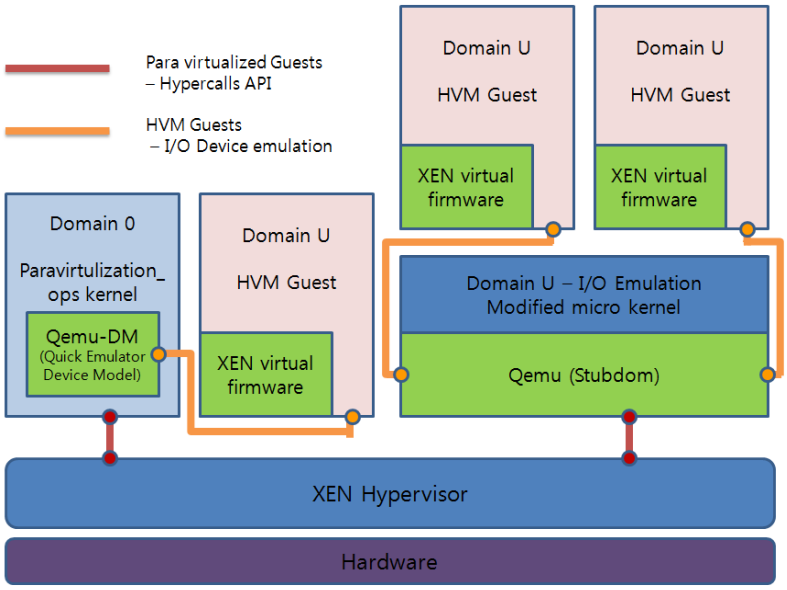


Figure 6. OpenXen Architecture and modules

OpenXen is an open source version of commercial Xen developed by Citrix Systems. OpenXen hypervisor should be installed on Linux system called Domain0 can control other guest OS called Domain1 ~ DomainU. The OpenXen hypervisor can control hardware directly even though it is installed on Linux system. However, sudden type of guest OS which should control BiOS directly like Microsoft's windows OS needs binary emulation called HVM (Hardware Virtual Machine). Qemu (Quick Emulator) is widely known binary translator for HVM and it has many variations according to its functional differences. [16][17]

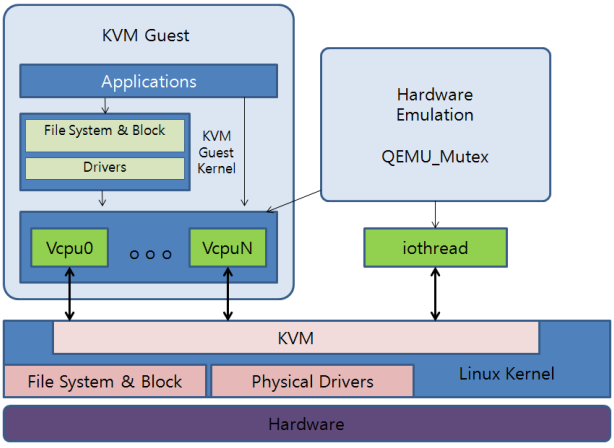


Figure 7. KVM hypervisor Architecture

KVM is open source hypervisor developed by Redhat, Inc. It is hybrid type hypervisor and controls guest OS at the same position level as that of Linux kernel. It also needs Qemu variations

to install Microsoft windows system. KVM has rather simple architecture by integrating advantages of traditional hypervisor and Kernel and is strong to handle multimedia and VDI.

2.2 Considerations on Multi-purpose VDI Protocol Design

Desktop sharing and control are main aspects of functions to enable virtualization. Virtual desktop running on a server can send screens to client device. There are several protocols used in desktop sharing. RDP, ICA, and RFB are widely deployed protocols. RDP is Microsoft's protocol and used in Microsoft RDP server and client. [18] ICA protocol is used in Citrix Systems'. [19] The two protocols are closed proprietary used in their own products and not published their structure. Otherwise, RFB protocol is opened its structure to the public. Many applications and products are using RFB as default protocol and it is also widely used. [20] The protocols stated above are 1st generation protocols and have limitation to implement desktop sharing.

Recently Redhat, Inc. takes an important role in Open Source Software by providing Redhat Linux OS, substituting legacy UNIX system. They published KVM hypervisor for cloud virtualization and desktop VDI client for effective VDI. Redhat's streaming protocol is classified as 2nd generation protocol, running on KVM hypervisor, is showing good performance with seamless playing on multimedia contents. However it is not supporting N-Screen devices and only working on its own cloud virtualization platform. Besides, current KVM hypervisor supports just one session between VDI client & server and is so limited.

2.3 Overall considerations on new platform concept

We consider new Smart TV platform having advanced system architecture. Firstly, it should have capability to connect numbers of N-Screen devices into one virtual OS guest residing on cloud virtualization server and users can see same screen via various N-Screen devices. Secondly, it should be able to use 2nd generation VDI protocols to support both document mode and streaming mode to cover office and individual requirements. It should also support smart device apps running with N-Screen features. Thirdly, the platform should be able to use N-Screen user's device as VDI client without client hardware dependency. There should be no request to prepare additional device for Smart TV. To use apps, if he who uses Android device, he could play Android apps on his Android device as he did. For iOS device user, he also could run iOS apps as he did. The user might get computing resources from any guests (i.e.: Windows, Linux, Android, etc.) of the virtualization server except the case of the closed OS like iOS which is not be able to be invited as a guest to the virtualization server.

Considering as a VDI client device, the user must utilize legacy desktop PC resources besides Android and iOS devices. With reflecting the features stated above, we design advanced flexible Smart TV platform & architecture to be able to use Smart TV system as Post PC.

3. NEW PLATFORM DESIGN AND ITS ARCHITECTURE

3.1 Functionality and coverage of the platform

To overcome the weakness & shortcoming of previous approaches, reflecting overall considerations on new platform stated above, we suggest flexible architecture concept of designing cooperative document work functionality, N-Screen capability, and multi-user resource sharing based on existing cloud virtualization and VDI architecture. Following requirements are covered by new platform proposal.

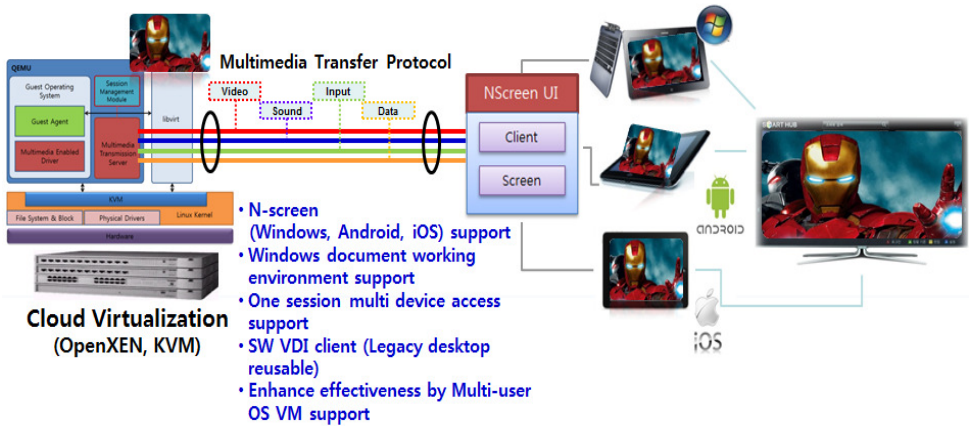


Figure 8. Module design diagram of new Smart TV platform

- Real-time multimedia support from VDI server for Smart TV server platform with Multimedia Transfer Protocol featuring 2nd generation VDI protocol
- Virtual Guest (VDI server for Smart TV) based desktop sharing & control performance providing document work functionality (office and individual work environment support)
- VDI S/W clients for N-Screen devices including legacy desktop PCs and thin client
- One server session with multi VDI streams to support group-watch or switching N-Screen devices
- Any network environment support using network tunneling server
-

3.2 Multimedia support VDI protocol and related module design

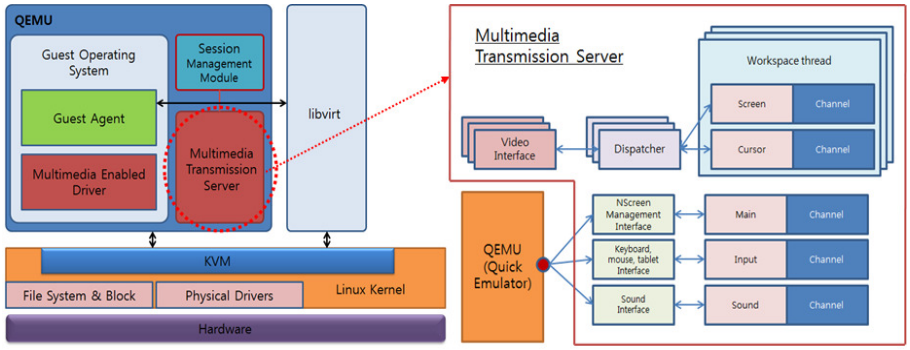


Figure 9. Multimedia Transfer Protocol diagram for the Smart TV Server and the Client

As we acknowledged from the previous approaches, there were cross-platform problems aroused from the provider of OS and hardware specification. Thus, new platform is basically designed to support OSS (Open Source Software) and to have flexibility and rather to be free from license issue. To solve the problem of OS-dependent cross-platform issue, our new Smart TV platform delivers full screen from the server to the client using VDI technology. Multimedia transmission server of the server side could handle this request of transferring steady high-resolution N-Screen delivery on both multimedia contents and document screen. The protocol is capable of delivering distinguished channels to carry screen delivery and access control separately. Its architecture is reducing conflicts and showing good performance especially on multimedia mode. Figure 9 shows that the structure of multimedia transmission server to handling each channel of delivery data and access & control signals.

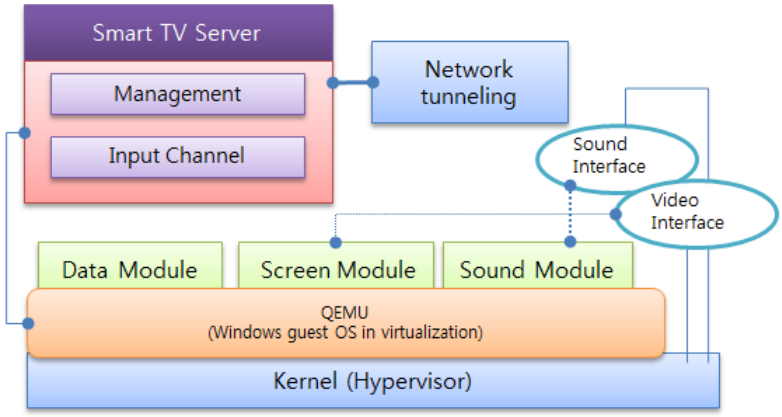


Figure 10. Smart TV server modules and stack structure

Even though the design was based on OSS, it is inevitable to support Microsoft Windows which is most commonly used and the user usually has a bundled license. To support MS Windows as a guest on the virtualization environment, there is a need to deploy the method of RDVH (Remote Desk Virtual machine Host) using Qemu (Quick emulator). Regarding Smart TV client, we consider legacy desktop PCs which are mostly using MS Windows OS and emerging smart devices running under Android OS or iOS.

Smart TV server consisted of several modules: the management module to process the request of the client by channels, the input channel to process keyboard and mouse value, the screen channel to transfer screens, the sound channel to process sounds, the data channel to process data between the server and the client, and networking module to process network handshaking and advanced pier to pier network tunnelling.

The sound module or the screen module need to communicate with hardware should interface via VGA driver or the sound driver installed in the OS kernel. Figure 10 shows Smart TV server modules and structure of the server channel by channel.

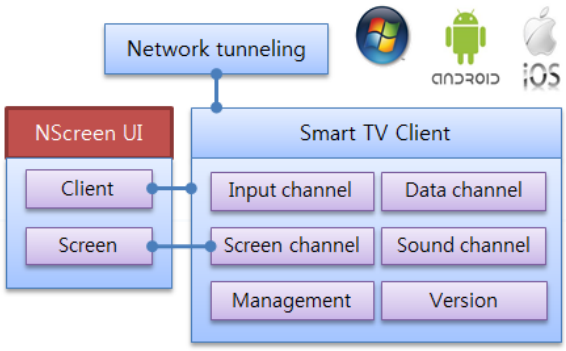


Figure 11. Modules and stack structure of the Client

The Smart TV client consisted of the almost same modules that have to cooperate with those of the server. The client consisted of the input channel, the sound channel, the screen channel, the data channel, and the network channel. N-Screen UI is designed to support N-Screen device and the client management module is to manage sessions and versions. Figure 11 shows the client module structure of channels and functions.

The data transfer protocol design for the server and the client are targeting to transfer each data using multi-channels and adjusting the requests of each other by handshaking traffic, reducing buffering, keeping steady packet transfer, and aiming stable communications. The user friendly protocol should be considered to deploy easily on the server and the client. It is also considered to design adding more smart device or OS in the future.

3.3 Screen Share to watch the same screen to N-Screen devices (Group watch)

The previous Smart TV approaches have not Screen Share function to watch the same screen with N-Screen devices. Following image shows the concept.

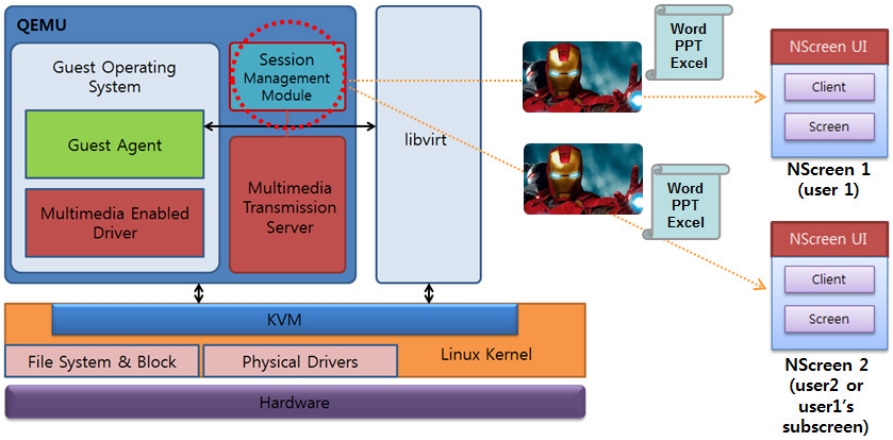


Figure 12. Smart TV to support N-Screen

The Session Management Module of the server could provide the function of Screen Share to send the screen on the main N-Screen display and many other sub screens of smart devices. It could be provided as a form of replicating stream to share the same handle.

3.4 N-Screen extended pack design and network tunnelling module

To utilize the user owned devices, it should include Android and iOS smart devices as basic client owned hardware. VDI client should be carefully designed because mobile OS like Android OS or iOS have limitations on multi process handling, memory handling, etc. Figure 13 shows Android modular stack architecture and virtual machine structure.

Considering N-Screen devices, even though mainly Android OS, iOS should be considered, Windows OS installed desktop PC should be included as good computing resource in customer side. Additionally new mobile OS might be added on the list in the future. It depends on the market needs and it should be also considered. Moreover, it should be reminded that the mobile OS periodically published new major and minor version much more frequently than that of desktop PC. To support N-Screen devices, it is also considered additional co-working features and functions: file transfer, system information, process information, screen capture, etc.

For seamless network connection, the network tunnelling server is inevitable for the clients residing on local private network. The network tunnelling server should be designed to process high stress of simultaneous session establishment requests. The server also should be considered active-active or active-standby backup server to support high availability.

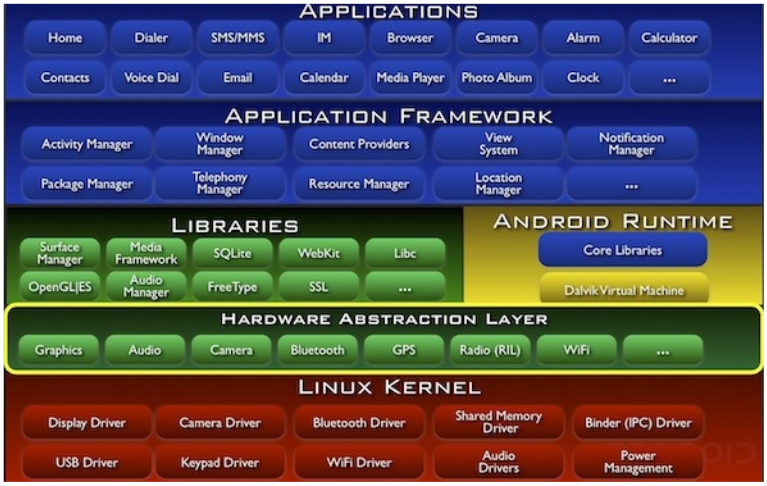


Figure 13. Android modular stack architecture

4. CONCLUSION & FURTHER CONSIDERATIONS

Reviewing architecture comparison focused on functions provided to Smart TV users, the new platform provides more flexibility compared to the previous approaches. New Smart TV platform is effective in that it consists of pure software based server and thin client to support N-Screen devices even including legacy desktop resources. It also provides both multimedia mode and document mode to support office work and individual needs. Following figure will show the benefits of proposed new Smart TV platform.

Figure1. Legacy Smart TV system, Android TV, iTV, and new platform

TV Type / Coverage	Legacy Smart TV	Android TV	iTV	New Platform (Virtualization)
Internet	O	O	O	O
Cloud server (Movie files)	O	O	O	O (Multimedia mode)
Thin client	X (Specific device)	Android only (limited)	iOS only (limited)	O (Any OS on his own)
Legacy desktop PC as the client	X	X	X	O (S/W installation)
apps	X	Android only (limited)	iOS only (limited)	O (On his own device)
Document work (PowerPoint, Excel, Word, etc.)	X	Limited	Limited	O (Windows guest, Document mode)

Regarding functions to be developed or dispatched according to the design above, the new platform should utilize generic functions provided by cloud virtualization technology and related open technologies. Following figures show the new platform's coverage compared to legacy desktop sharing and pure hypervisor based open software platform.

Table 2. Screen share functionality comparison

Functions/ Types	Legacy Desktop sharing	Hypervisor based Open S/W	New Platform (Hypervisor based)
Desktop sharing	O	O	O
Multi-channel protocol & management	X	O (VDI protocol)	O (development or alteration needed)
Video streaming mode	X	O	O(same as above)
1:n multiuser view (N-Screen)	O	X	O(same as above)
Local network support	X	X	O(same as above)

As a further discussion, new platform should enhance effectiveness by deploying a multi-user OS for virtualization guest. Up to now, cloud virtualization server allows just one session between the server and a client. Though new platform design revises it to enable 1: n sessions and support N-Screen, original concept is based on 1:1 VDI concept. To support multi-user OS in virtualization server, it seemed that there need lot of efforts to modify it. However, if many users can access to a multi-user OS guest in VDI server for Smart TV, it becomes very effective approach to upgrade the capability of the platform as shown Figure 14.

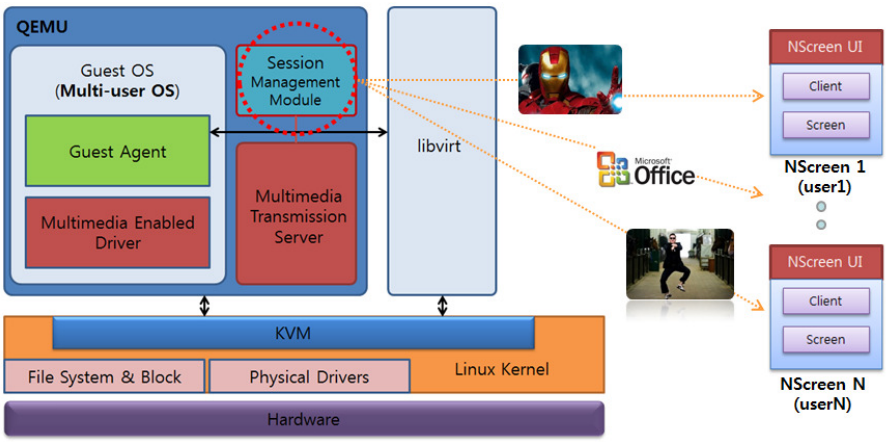


Figure 14. Consideration on the multi-user OS guest support

REFERENCES

[1] Steve Kovach, "What Is A Smart TV?", http://www.businessinsider.com/what-is-a-smart-tv-2010-12, Businessinsider.com, December 8, 2010

[2] Adrian Kingsley-Hughes,"PC, post-PC... what next?", http://www.zdnet.com/pc-post-pc-what-next-7000018071/, zdnet.com, July 15, 2013

[3] "Collaboration in the PostPC era", http://www.cisco.com/c/dam/en/us/products/collateral/unified-communications/jabber-android/cisco-collab-at-a-glance.pdf, Cisco.com, 2012

[4] Tobin, A., "A Sky in the Cloud: The UltraViolet initiative has given a boost to the notion of cloud-based digital lockers as the route to secure multiscreen on-demand availability. But does the concept make sense for pay TV operators?", Digital TV Europe, Vol. No. 306, pp. 24-29, Informa Telecoms & Media, 2012

[5] JA Kim, DH Kim, "A Study on the Screen Evolution and Expansion of the Concept", The Journal of Korea Institute of Next Generation Computing, Vol.8 No.2, pp.87-98, Korea Institute of Next Generation Computing, 2012

[6] Samsung TV portal, http://www.samsung.com/global/article/articleList.do?articleMode=category&lndSiteCode=uk&selectedCtgry=3&selectedCtgryOrder=1&page=1, Samsung.com, 2013

[7] EJ. Choi, HW. Song, JW. Lee, CS. Bae, "Web-based Personal application Managements in Personal Cloud Computing Environments", The Journal of Korea Institute of Next Generation Computing, Vol.10 No.1, pp. 65-73, Korea Institute of Next Generation Computing, 2014

[8] Google TV, "How it works?", http://www.google.com/tv/features.html, 2014

[9] Chris Smith, "More evidence suggests Apple's 'iTV' plans are real", http://bgr.com/2014/03/06/apple-itv-release-date/, BRGMedia, LLC, 2014

[10] GW. Kim, WJ. Lee, CH. Jeon, "Virtualization technology for cloud computing", KSCI Review, v.18, no.1, pp.25-33, KSCI, 2010

[11] Tsai, H. Y. , Siebenhaar, M. , Miede, A. , Huang, Y. , Steinmetz, R.,"Threat as a Service?: Virtualization's Impact on Cloud Security", IT professional ,vol.14 no.1, IEEE, 2012

[12] Xing, Y. , Zhan, Y., "Virtualization and Cloud Computing", LECTURE NOTES IN ELECTRICAL ENGINEERING Vol.143, pp305-312, Springer Science + Business Media, 2012

[13] Hudic, A. , Weippl, E.,"Private Cloud Computing: Consolidation, Virtualization, and Service-Oriented Infrastructure", Computers & security Vol.31 No.4, Elsevier Science B.V., Amsterdam, 2012

[14] Yang, C.-T., Tseng, C.-H., Chou, K.-Y., Tsaur, S.-C., Hsu, C.-H., Chen, S.-C., A Xen-Based Paravirtualization System toward Efficient High Performance Computing Environments, Lecture Notes in Computer Science, Vol. No. 6083, pp126-135, SPRINGER-VERLAG, 2010

[15] "KVM and open virtualization: Who's using it, how and why?", IBM Systems and Technology, Thought Leadership White Paper, IBM, May 2013

[16] "Xen Architecture", http://en.wikipedia.org/wiki/Xen, Wikipedia.org, October 2012

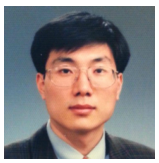
[17] "QEMU", http://en.wikipedia.org/wiki/QEMU, Wikipegia.org, September 2011

[18] "Remote Desktop Protocol", http://en.wikipedia.org/wiki/Remote_Desktop_Protocol, Wikipedia.org, May 2014

[19] "Independent Computing Architecture", http://en.wikipedia.org/wiki/Independent_Computing_Architecture, Wikipedia.org, May 2014

[20] "RFB protocol", http://en.wikipedia.org/wiki/RFB_protocol, Wikipedia.org, March 2014

Authors



JuByoung Oh (Mr.)
CEO & Chairman, Koino, Inc.
Senior Researcher, ETRI
Ph.D. Completed doctoral course in Computer Science Engineering, Chungnam National University

Ohseok Kwon (Mr.)
Professor, Computer Science Engineering Department, Chungnam National University
M.EE, KAIST
B.EE, Seoul National University

INTENTIONAL BLANK

PHONETIC CLASSIFICATION BY ADAPTIVE NETWORK BASED FUZZY INFERENCE SYSTEM AND SUBTRACTIVE CLUSTERING

Samiya Silarbi, Bendahmane Abderrahmane and Abdelkader Benyettou

Faculty of mathematical and computer science, department of Computer Science, University of Sciences and Technology Oran USTO-MB, Algeria.
samiya.silarbi@gmail.com, abder.bendahmane @gmail.com, a_benyettou@yahoo.fr

ABSTRACT

This paper presents the application of Adaptive Network Based Fuzzy Inference System ANFIS on speech recognition. The primary tasks of fuzzy modeling are structure identification and parameter optimization, the former determines the numbers of membership functions and fuzzy if-then rules while the latter identifies a feasible set of parameters under the given structure. However, the increase of input dimension, rule numbers will have an exponential growth and there will cause problem of "rule disaster". Thus, determination of an appropriate structure becomes an important issue where subtractive clustering is applied to define an optimal initial structure and obtain small number of rules. The appropriate learning algorithm is performed on TIMIT speech database supervised type, a pre-processing of the acoustic signal and extracting the coefficients MFCCs parameters relevant to the recognition system. Finally, hybrid learning combines the gradient decent and least square estimation LSE of parameters network. The results obtained show the effectiveness of the method in terms of recognition rate and number of fuzzy rules generated.

KEYWORDS

Phoneme, recognition, ANFIS, subtractive clustering.

1. INTRODUCTION

Automatic Speech Recognition has achieved substantial success in the past few decades but more studies are needed because none of the current methods are fast and precise enough to be comparable with human recognition abilities [1,2]. Many algorithms and schemes based on different mathematical paradigms have been proposed in an attempt to improve recognition rates [3,4,5,6].

Neuro-fuzzy modeling is a combination of fuzzy logic and neural network that takes advantage of both approaches, process imprecise or vague data by fuzzy logic [7] and at the same time by introducing learning through neural network. Several architectures have been proposed depending on the type of rule they include Mamdani or Sugeno [8] [9] one of the most influential fuzzy models has been proposed by Robert Jang in [10] called Adaptive Network Based Fuzzy Inference System ANFIS. The rule base of this model contains the fuzzy if-then rule of Takagi and Sugeno's type in which consequent parts are linear functions of inputs instead of fuzzy sets, reducing the number of required fuzzy rules.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 187–196, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4918

The identification of fuzzy model consists of two major phases: structure identification and parameter optimization. The first phase is the determination of number of fuzzy if-then rules and membership functions of the premise fuzzy sets while the second phase is the tuning of the parameter values of the fuzzy model [11]. However, there will be two problems if we directly use the traditional Takagi Sugeno model for the speech recognition system. The first one is the network reasoning will fail if the input dimension is too large. The second problem is that with the increase of input dimension, rule numbers will have an exponential growth and cause "rule disaster". Thus, determination of an appropriate structure becomes an important issue, then clustering techniques are applied to solve this problem. [12] [13] [14].

This paper present a neural fuzzy system ANFIS for speech recognition. The appropriate learning algorithm is performed on TIMIT speech database supervised type, a pre-processing of the acoustic signal and extracting the coefficients MFCCs parameters relevant to the recognition system. Subtractive clustering is applied in order to define an optimal structure and obtain small number of rules, then learning of parameters network by hybrid learning which combine the gradient decent and least square estimation LSE.

The paper is organized as follows: the section 2 reviews the literature research work, in section 3 details the structure and Principe of learning of ANFIS, while section 4 describes subtractive clustering; experimental results and discussion were detailed in section 5. Section 6 concludes the paper.

2. RELATED WORK

The ANFIS has the advantage of good applicability because it can be interpreted as local linearization modeling, and even as conventional linear techniques for both state estimation and state control which are directly applicable. This adaptive network has good ability and performance in system identification, prediction and control has been applied in many different systems. Since there are not many research works used ANFIS in speech recognition, it is necessary to carry on the exploration and the thorough research on it. ANFIS was used for Speaker verification [15] using combinational features of MFCC; Linear Prediction Coefficients LPC and the first five formants; Recognition of discrete words [16], Speech emotion verification [17] based on MFCC for real time application. In [18] ANFIS was performed to reduce noise and enhance speech, also for the recognition of isolated digits with speaker-independent [19]. Speaker, language and word recognition was completed by ANFIS [20], furthermore for caller behavior classification [21]. All of these works had used clustering techniques to determine the structure of ANFIS. Another work: an automated gender classification is performed by ANFIS [22].

3. ADAPTIVE NETWORK BASED FUZZY INFERENCE SYSTEM ANFIS

3.1. The Concept and Structure

ANFIS proposed by Jang in 1993 multi-layered neural network which connections are not weighted or all weights equal 1**[10]**, is alternate method which combines the advantages of two intelligent approaches neural network and fuzzy logic to allow good reasoning in quantity and quality. A network obtained has an excellent ability of training by means of neural networks and linguistic interpretation of variables via fuzzy logic. The both of them encode the information in parallel and distribute architecture in a numerical framework. ANFIS implement a first order Sugeno style fuzzy system; it applies the rule of TSK Takagi Sugeno and Kang form in its architecture.

Rule: if x is A1 and y is B1 then $f(x) = px + qy + r$

Where x and y are the inputs, A and B are the fuzzy sets, f are the output, p, q and r are the design parameters that determined during the training process. ANFIS is composed of two parts is the first part is the antecedent and the second part is the conclusion, which are connected to each other by rules in network form. Five layers are used to construct this network. Each layer contains several node sits structure shows in figure 1.

layer1: executes a fuzzification process which denotes membership functions (MFs) to each input. In this paper we choose Gaussian functions as membership functions:

$$O_i^1 = \mu_{Ai} = \exp\left(\frac{-(x-c)^2}{\sigma^2}\right) \tag{1}$$

layer2: executes the fuzzy AND of antecedents part of the fuzzy rules

$$o_i^2 = w_i = \mu_{Ai}(x_1) \times \mu_{Bi}(x_2), i = 1,2,3,4 \tag{2}$$

layer3: normalizes the MFs

$$o_i^3 = \overline{w_i} = \frac{w_i}{\sum_{j=1}^{4} w_j}, i = 1,2,3,4 \tag{3}$$

layer4: executes the conclusion part of fuzzy rules

$$O_i^4 = \overline{w_i} y_i = \overline{w_i}\left(\alpha_1^i x_i + \alpha_2^i x_2 + \alpha_3^i\right), i = 1,2,3,4. \tag{4}$$

layer5: computes the output of fuzzy system by summing up the outputs of the fourth layer which is the defuzzification process.

$$O_i^5 = overll\_output = \sum_{i=1}^{4} \overline{w_i} y_i = \frac{\sum_{i=1}^{4} w_i y_i}{\sum_{i=1}^{4} w_i} \tag{5}$$

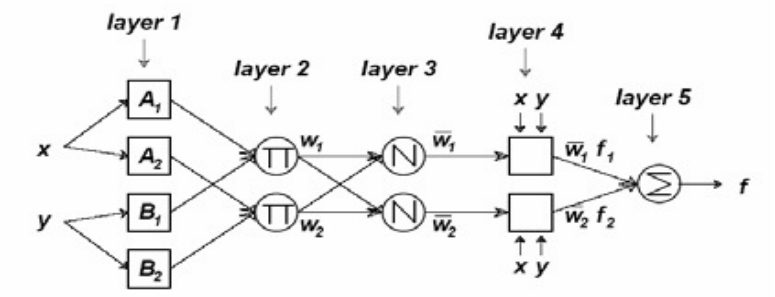


Figure 1. ANFIS architecture.

Circles in ANFIS represent fixed nodes that predefined operators to their inputs and no other parameters but the input participate in their calculations. While square are the representative for adaptive nodes that affected by internal parameters.

3.2. Learning Algorithm

The parameters to be tuned in an ANFIS are the membership function parameters of each input, the consequents parameters also number of rules.

$$nbr\_rule = m^n \tag{6}$$

Where n is number of inputs and m number of membership functions by input and they generate all the possible rules.

Two steps of training are necessary which are: Structure learning which allows to determinate the appropriate structure of network, that is, the best partitioning of the input space (number of membership functions for each input, number of rules). And parametric learning carried out to adjust the membership functions and consequents parameters. In most systems the structure is fixed a priori by experts. In our work we combine both of learning sequentially

The subsequent to the development of ANFIS approach, a number of methods have been proposed for learning rules and for obtaining an optimal set of rules. For example, Mascioli et al **[23]** have proposed to merge Min-Max and ANFIS model to obtain neuro-fuzzy network and determine optimal set of fuzzy rules. Jang and Mizutani **[24]** have presented application of Lavenberg-Marquardt method, which is essentially a nonlinear least-squares technique, for learning in ANFIS network. In another paper, Jang **[25]** has presented a scheme for input selection and **[26]** used Kohonen's map to training.

Four methods have been proposed by Jang **[11]** for update the parameters of ANFIS:

- All parameters are update by only gradient decent.
- In first the consequents parameters are obtained by application of least square estimation LSE only once and then the gradient decent update all parameters.
- Sequential LSE that is using extended Kalman filter to update all parameters.
- Hybrid learning: which combine the gradient decent and LSE to find a feasible set of antecedents and consequents parameters.

The most common training algorithm is the hybrid learning. this algorithm is carried out in two steps: forward pass and backward pass, Once all the parameters are initialized, in forward pass, functional signals go forward till fourth layer and the consequents parameters are identified by LSE. After identifying consequents parameters, the functional signals keep going forward until the error measure is calculated. In the backward pass, the error rates propagate backward and the premise parameters are updated by gradient decent.

The function to be minimized is Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\left(\frac{1}{N} \sum_{i=1}^{N} (d_i - o_i)^2 \right)} \tag{7}$$

Where d_i is the desired output and O_i is the ANFIS output for the ith sample from training data and N is the number of training samples.

4. SUBTRACTIVE CLUSTERING

Clustering is division of data into groups of similar objects. Each group (cluster) consist of objects that are similar between themselves and dissimilar to objects of others groups. From the machine learning perspective, clustering can be viewed as unsupervised learning of concepts **[27][28]**.

1. Compute the initial potential value for each data point x_i is defined as:

$$P_i = \sum_{j=1}^{n} \exp\left[- \frac{\left\| x_i - x_j \right\|^2}{\left(ra / 2 \right)^2} \right] \tag{8}$$

Where ra is a positive constant representing a neighborhood radius. Therefore, a point would have a height potential value if it has more neighbor points close to itself.

2. The point with the highest potential value is selected as le first cluster center: First cluster center x_{c1} is chosen as the point having the largest density value D_{c1}

$$P^{(1)*} = \max_i \left(P^{(1)}\left(x^i \right) \right) \tag{9}$$

3. The potential value of each data point x_i is reduced as follows:

$$P_i = P_i - P_{c1} \exp\left[- \frac{\left\| x_i - x_{c1} \right\|^2}{\left(rb / 2 \right)^2} \right] \tag{10}$$

Where ($rb = \beta * ra$) is a positive constant represent the radius of the neighborhood for which significant potential reduction will occur.

β is a parameter called as squash factor, which is multiply by radius values to determine the neighboring clusters within which the existence of other cluster centers are discouraged.

1. After revising the density function, the next cluster center is selected as the point having the greatest potential value. This process is repeated to generate the cluster centers until maximum potential value in the current iteration is equal or less than the threshold.

5. STRUCTURE LEARNING ALGORITHM FOR T-S FUZZY NEURAL NETWORK

ANFIS offers three approaches to identify cluster namely grid partitioning, subtractive clustering and fuzzy c-means clustering. Grid partitioning approach is useful if the number of features is no more than 6 or 7. If the number of features is too high then this method will cannot be used as the memory requirement will be insufficient when using MATLAB. For fuzzy c-means method, the number of clusters for the dataset needs to be specified. Since no prior knowledge on the number of clusters is available, subtractive clustering will be ideal.

The radius of the clusters will be used as the basis of the cluster formation. A range of radius from 0.0 to 2.2 has been chosen in order to produce optimum number of clusters for the initial fuzzy inference system (FIS). The initial FIS then was fed to the ANFIS for refining the FIS so that it will tune the supervised learning iteratively with more detailed rules generated.

Each cluster center represents a fuzzy if-then-rule. The nth column of cth cluster center is assumed to be the mean value (c_{in}) of the associated Gaussian membership function defined for cth fuzzy set of nth input variable. Then, the standard deviation (a_{in}) of the above mentioned Gaussian functions are calculated as below:

$$a_{in=re_n}\left(\frac{\max(x^n)-\min(x^n)}{\sqrt{s}}\right) \quad (11)$$

Therefore the cluster centers and squash factors may be viewed as parameters, which the number of fuzzy rules of the initial FIS depend on, before the rule base parameters of the FIS is tuned by ANNs in ANFIS.

6. EXPERIMENTAL AND RESULTS

6.1. TIMIT Database

A reduced subset of TIMIT **[29]** -Phonetic Continuous Speech Corpus (TIMIT – Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)) used in our work, which is the abridged version of the complete testing set, consists of 192 utterances, 8 from each of 24 speakers (2 males and 1 female from each dialect region). In general, phonemes can grouped into categories based on distinctive features that indicate a similarity in articulatory, acoustic and perceptual. The major classes of phonetic TIMIT database are: 6 vowels {/ah/, /aw/, /ax/, /ax-h/, /uh/, /uw/}, 6 fricatives {/dh/, /f/, /sh/, /v/, /z/, /zh/} and 6 plosives {/b/, /d/, /g/, /p/, /q/, /t/}.

6.2. Coding Mel-Frequency Cepstral Coefficients MFCC

Before any calculation, it is necessary to effect some operations to shape the speech signal. The signal is first filtered and then sampled at a given frequency (16 KHZ). Pre-emphasis is carried out to raise the high frequencies. Then, the signal is segmented into frames, each frame has a fixed number N=25ms of speech samples (period in which the speech signal can be considered as stationary).Treating small fragments of signal leads to filtering problems (edge effects). In order to reduce edge effects we use weights windows (Hamming window) these are functions that are applied to set of samples in the window of the original signal.

After signal shaping a discrete Fourier transform DFT particularly fast Fourier transform FFT is applied to pass in the frequency domain and for extracting the spectrum signal. Then, the filtering is performed by multiplying the spectrum obtained by filters (triangular filters). It is possible to employ the output of the filter bank as input to the recognition system. However, other factors derived from the outputs of a bank of filters are more discriminative, more robust and less correlated. It is from the cepstral coefficients derived of the filter bank outputs distributed linearly on the Mel scale, these are the Mel-frequency Cepstral coefficients MFCC Each window provides 12 MFCC coefficients and the corresponding residual energy.

6.3. Results

The corpus that we have used in the classification of phonemes consists of 6 vowels, 6 fricatives and 6 plosives: 31514 instances for learning and 12055 instances for test. Applying subtractive clustering to determinate the initial structure of ANFIS, then Hybrid learning to find a feasible set of antecedents and consequents parameters.

We applied ANFIS on the classification of phonemes by varying the radius of the clustering between [0.0- 2.5], results are summarized in the following tables:

With a radius between [0-1.4] we achieved a recognition rate of 100 % but the number of generated rules was too big, so that exceeds 200 rules which has resulted a large runtime time and especially for real-time application, and this for the three classes of phonemes used.

For vowels we have obtained the best recognition rate of 100 % and 6 rules with a radius of 2.0 and 1.9, we have reduced the number of rules from 6 to 4 with a radius of 2.1 and achieved recognition rate of 83 % (table 1).

For fricatives with a radius of 1.4 we have obtained the best recognition rate that reaches 100 % and 5 fuzzy rules, we could reduce the number of rules to 4 with a recognition rate of 80.66 % (table 2).

For plosives we reached 100 % recognition rate and 5 rules with a radius of 1.4 and 1.5 beyond 1.5 we had 3 rules but reduction of recognition (table 3).

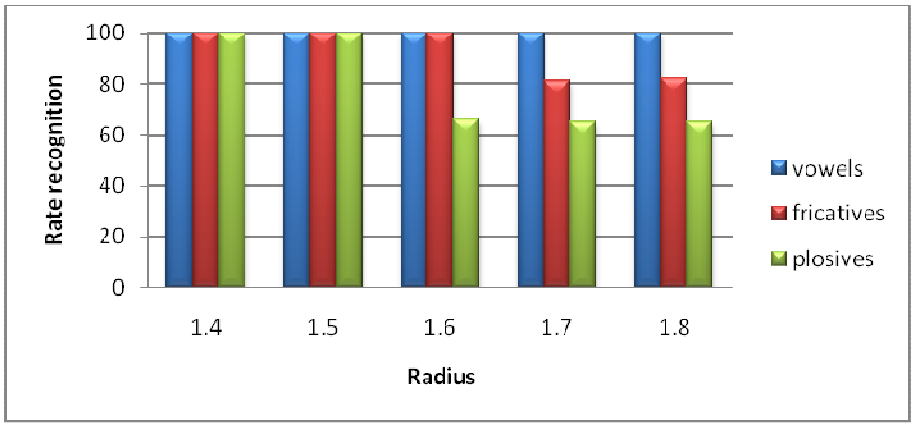


Figure 2. Rate recognition depending of radius

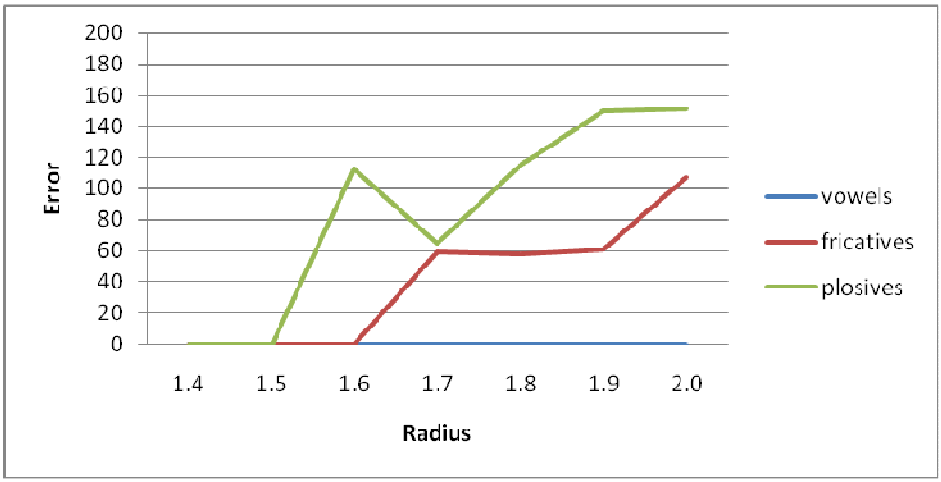


Figure 3. Error depending of radius

Table 1. Vowels.

Radius	Rate-recognition%		error		Number-rules	
	train	test	train	test	train	test
1.4	100	100	0	0	37	34
1.5	100	100	0	0	22	23
1.6	100	100	0	0	17	15
1.7	100	100	0	0	11	10
1.8	100	100	0	0	9	8
1.9	100	100	0	0	8	6
2.0	100	100	0	0	7	5
2.1	100	83	0	53	7	4
2.2	78	79	66	62	4	4

Table 2. Fricatives.

Radius	Rate-recognition%		error		Number-rules	
	train	test	Train	test	train	test
1.4	100	100	0	0	5	5
1.5	64.33	100	120	0	3	6
1.6	58.66	100	139	0	3	5
1.7	59	80.66	137	59	30	4
1.8	58	81	138	58	3	4

Table 3. Plosives.

Radius	Rate-recognition%		error		Number-rules	
	train	test	train	test	train	test
1.4	100	100	0	0	6	5
1.5	100	100	0	0	8	5
1.6	100	65.33	0	113	5	3
1.7	89	65	34	65	4	3
1.8	89	65	34	115	3	2

A large number of rules are generated by reducing the radius of clustering so most of data are correctly classified but it takes significant time. Contrariwise, by increasing the radius of clustering we had very few rules so many alternatives are not considered by the fuzzy rules, then a lot of data are not correctly classified.

We could see that there was a compromise between the recognition rate and the number of rules generated, with a radius between [1.5-1.8] we were able to achieve perfection on the recognition rate and number of generated rules which engender a very reasonable computation time.

7. CONCLUSION AND FUTURE WORK

In this paper, we have applied adaptive network fuzzy inference system for phonemes recognition. First learning of the network structure by subtractive clustering, in order to define an optimal structure and obtain small number of rules, then learning of parameters network by

hybrid learning which combine the gradient decent and least square estimation LSE to find a feasible set of antecedents and consequents parameters.

The appropriate learning algorithm is performed on TIMIT speech database supervised type, a pre-processing of the acoustic signal and extracting the coefficients MFCCs parameters relevant to the recognition system. Finally, hybrid learning combines the gradient decent and least square estimation LSE of parameters network. The results obtained show the effectiveness of the method in terms of recognition rate and number of fuzzy rules generated.

For future work, consider this adaptive network for whole dataset TIMIT, and improving learning algorithm by tuning parameters of ANFIS by means of evolutionary algorithms.

REFERENCES

[1] R. Reddy (2001), Spoken Language Processing: A guide to Theory, Algorithm, And System Development, Prentice-Hall, New Jersey.

[2] X. He, L. Deng (2008), Discriminative Learning for Speech Recognition: Theory and practice, Morgan & Claypool.

[3] Cole, Ron, Hirschman, Lynette, Atlas, Les, Beckman, Mary, Biermann, Alan, Bush, Marcia, et al (1995) The challenge of spoken language systems: Research directions for the nineties. IEEE Transactions on Speech and Audio Processing, vol 3(issue 1).

[4] Scofield, M. C (1991) Neural networks and speech processing. Amsterdam: Kluwer Academic.

[5] Yallop, C. C. (1990) An introduction to phonetics and phonology. Cambridge, MA: Blackwell.

[6] Carla Lopes and Fernando Perdigão. chapter 14 (2011) Phone Recognition on the TIMIT Database. Book Speech Technologies Edited by Ivo Ipsic, ISBN 978-953-307-996-7, 432 pages, Publisher InTech.

[7] George Bojadziev,Maria Bojadziev (1995) Advances in fuzzy systems applications and theory, vol 5 book Fuzzy Sets, Fuzzy Logic, Applications, World Scientific.

[8] Ajith Abraham (2001) "Neuro Fuzzy Systems: State-of-the-Art Modeling", Techniques in Proceedings of 6th International Work-Conference on Artificial and Natural Neural Networks, IWANN 2001 Granada, Spain, June 13–15, Springer Verlag Germany, vol 2084, pp 269-276.

[9] B. Kosko (1991) "Neural Networks and Fuzzy Systems A Dynamic Systems Approach", Prentice Hall.

[10] J.S.R. Jang (1993) "ANFIS: Adaptive Network Based Fuzzy Inference Systems", IEEE Trans, Syst. Man Cybernet. vol 23, No 3, pp. 665-685.

[11] J.S.R. Jang, C.T.Sun and E.Mizutani (1997) Neuro-fuuzy and soft computing: a computational approach to learning and machine intelligence, London: prentice-Hall international, USA.

[12] Agus Priyono, Muhammad Ridwan, Ahmad Jais Alias, Riza Atiq O. K. Rahmat, Azmi Hassan & Mohd. Alauddin Mohd. Ali (2005) "Generation of fuzzy rules with subtractive clustering". Jurnal Teknologi, 43(D). Universiti Teknologi Malaysia, pp. 143–153.

[13] Chuen-Tsai Sun (1994) "Rule-Base Structure Identification in an Adaptive-Network-Based Fuzzy Inference System". IEEE Transaction on Fuzzy Systems, vol. 2, no. 1. pp. 64 – 73.

[14] Ching-Chang Wong and Chia-Chong Chen (1999) "A Hybrid Clustering and Gradient Descent Approach for Fuzzy Modeling". IEEE Transactions on systems, man, and cybernetics—part b: cybernetics, vol. 29, no. 6. pp. 686 – 693.

[15] V .Srihari, R.Karthik and R.Anitha and S.D.Suganthi (2010) "Speaker verification using combinational features and adaptive neuro-fuzzy inference systems", IIMT'10 December 28-30, Allahabad, UP, India, pp. 98-103.

[16] N.Helmi, B.H.Helmi (2008)"Speech recognition with fuzzy neural network for discrete words", ICNC Fourth International Conference on Natural Computation. vol 07, pp. 265 – 269.

[17] N.Kamaruddin, A.Wahab (2008) "Speech emotion verification system (sevs) based on mfcc for real time application", Intelligent Environments, 2008 IET 4th International Conference on, Seattle, pp. 1-7.

[18] Reem Sabah & Raja N. Ainon (2009) "Isolated Digit Speech Recognition in Malay Language using Neuro-Fuzzy Approach", Third Asia International Conference on Modelling & Simulation AMS, , Bali Asia, pp. 336 – 340.

[19] Jasmin Thevaril and H.K.Kwan (2005) "Speech Enhancement using Adaptive Neuro-Fuzzy Filtering" in Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems ISPACS, December 13-16, pp753 – 756.

[20] Bipul Pandey, Alok Ranjan, Rajeev Kumar and Anupam Shukla (2010) "Multilingual Speaker Recognition Using ANFIS". in 2nd International Conference on Signal Processing Systems (ICSPS) vol 3, V3-714 - V3-718.

[21] Pretesh. B. Patel, Tshilidzi Marwala (2011) "Adaptive Neuro Fuzzy Inference System, Neural Network and Support Vector Machine for caller behavior classification", 10th International Conference on Machine Learning and Applications ICMLA, 18-21 December, vol 1, pp 298 - 303.

[22] Sachin Lakra, Juhi Singh and Arun Kumar Singh (2013) "Automated Pitch-Based Gender Recognition using an Adaptive Neuro-Fuzzy Inference System", International Conference on Intelligent Systems and Signal Processing (ISSP), pp 82 – 86.

[23] Gujarat.Manish Kumar, Devendra P. Garg (2004) "Intelligent Learning of Fuzzy Logic Controllers via Neural Network and Genetic Algorithm", Proceedings of 2004 JUSFA Japan – USA Symposium on Flexible Automation. Denver, Colorado, pp. 1-8.

[24] Mascioli, F.M., Varazi, G.M. and Martinelli, G (1997) "Constructive Algorithm for Neuro-Fuzzy Networks", Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, Vol. 1, pp. 459 -464.

[25] Jang, J.-S. R., and Mizutani, E (1996) "Levenberg-Marquardt Method for ANFIS Learning", Biennial Conference of the North American Fuzzy Information Processing Society, pp. 87 -91.

[26] Jang, J.-S.R. (1996) "Input Selection for ANFIS Learning", Proceedings of the Fifth IEEE International Conference on Fuzzy Systems, Vol. 2, pp. 1493 -1499.

[27] J.Han, M.Kamber (2000) "Data Mining: Concepts and Techniques" chapter 8. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers.

[28] V.Estivill-Castro, J.Yang (2000) "A Fast and robust general purpose clustering algorithm". Pacific Rim International Conference on Artificial Intelligence, pp. 208-218.

[29] Zue, V.; Seneff, S. & Glass J (1990) "Speech database development at MIT: TIMIT and beyond", Speech Communication, Vol. 9, No. 4, pp. 351-356.

A NEW APPROACH OF CONCURRENT CALL HANDLING PROCEDURE IN MOBILE NETWORKS

P. K. Guha Thakurta[1] Misha hungyo[2] Jahnavi Katikitala[3] and Darakshan Anwar[4]

[1,2,3,4] Department of Computer Science and Engineering,
National Institute of Technology, Durgapur, WestBengal
[1]parag.nitdgp@gmail.com,
[2]mkonghar@gmail.com,
[3]jahnavi6392@gmail.com,
[4]dara.singh09876@gmail.com

ABSTRACT

A new approach for handling concurrent call requests by a number of senders in mobile cellular network is proposed in this paper. The concurrent access is resolved with the introduction of semaphore concept. The several factors are identified to establish a priority factor (PF) for the sender node. Based on this PF value, the right sender is selected by the receiver in case of concurrent requests. This selection algorithm executes in linear time. The effectiveness of the proposed model is analyzed with the introduction of progress graph.

KEYWORDS

Concurrent call requests, semaphore, mobile networks, progress graph

1. INTRODUCTION

In mobile cellular networks, mobile nodes communicate with each other using multi-hop links. This structure is stationary because there are base stations (nodes) in every cell. Each node in the network has call forwarding capability to other nodes [5]. Till date, various routing strategies have been designed to address the problem of finding routing path with efficient congestion control technique. Simultaneously, it needs a more efficient methodology to increase throughput and reduce network latency at the same time. To provide the efficient routing strategy, the nodes are grouped into manageable clusters based on different parameters and the Quality of Service (*QoS*) availability of each node. Here, the cluster heads (*CHs*) play the role of local coordinators and they maintain the *QoS* values of all cluster members. Using this information, a *CH* can forward the calls to the corresponding destinations [4]. Thus, the network design problem associated with this is to find a least cost or a maximum revenue network, reducing the redundant admitted calls.

Once the clusters are formed and maintained, these can be used to handle incoming call requests. When a mobile node sends a call request, the call is forwarded through the path as: *CS→BSS→LeaderS→LeaderR→ BSR→CR* [1], where *CS , BSS* and *LeaderS* denote the sender's node, its corresponding *BS* and the *CH* of that cluster respectively. Similar notations are used for

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 197–204, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4919

the receiver part, i.e., *CR, BSR* and *LeaderR* respectively. The concurrent call handling procedure for multiple numbers of sender nodes under such scenario is not discussed. Moreover, the number of unsuccessful attempts by a sender node is not taken into consideration in this procedure and so, no priority is assigned on this basis. This may lead to indefinite waiting time for such node.

In this paper, a new approach for handling concurrent call requests by the multiple numbers of sender nodes in mobile cellular networks is proposed. The concurrent access is resolved with the introduction of semaphore [2] concept. The priority factor (*PF*) is calculated by the receiver in case of multiple numbers of senders. Here, PF is dependent on the available bandwidth (*BW*) to handle call request, timestamp (t_s), and repeated request factor (*RRF*) of the sender node. With the help of this *RRF* value, the number of unsuccessful attempts by a sender node is defined. Thus the node with the highest *PF* is selected as the right one. In order to avoid indefinite waiting time for the selection by receiver node, ts and *RRF* values of the specific sender whose requests had been processed already, are set to zero. With the introduction of *PF*, the proposed model in this work executes in linear time. In addition, the effectiveness of the proposed model is analyzed using progress graph.

The remainder of the paper is organized as follows. In section II, the proposed model is described. Next, the advantage of this work is concluded in section III.

2. PROPOSED MODEL

The model proposed in this work obeys the following system model.

2.1. System Model and assumptions

Suppose, there are three clusters of nodes shown in Fig. 1. Two of them represent the sender's clusters among three, whereas the remaining one denotes the receiver's cluster. The sender nodes and are trying concurrently to connect with the receiver through their respective CHs.

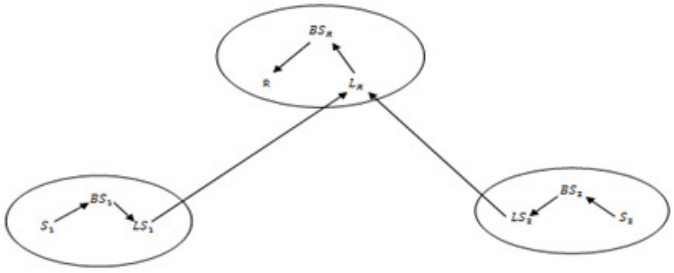


Fig. 1: Concurrent call requests from multiple sender's to a receiver

Now, the receiver node has a responsibility to select one of them having the highest priority factor (PF). This PF is dependent on three following factors.

(a) Bandwidth (*BW*): It is the amount of available *BW* to handle the call request. Generally, it depends on the underlying hardware architecture and the network operating system used by the sender node.

(b) Timestamp: The t_s of any call request is determined by the call submission time in the system and it is measured as the time recorded by the system clock. The request with least timestamp would become as the oldest call.

(c) Repeated Request Factor (*RRF*): It is defined as the number of unsuccessful connection establishment by the sender node. Initially, it is set to 0 and is incremented by 1 after each unsuccessful attempt. Whenever the receiver selects the specific sender node for the connection establishment, then *RRF* is again set as 0 for that sender. The node having greater *RRF* value is considered as higher priority node. As available bandwidth limits the number of call requests that can be processed, so *PF* is directly proportional to bandwidth *BW*. With the help of *RRF* factor, a preference is given to a sender node which is calling repeatedly. So, a node with greater *RRF* value has a greater *PF* value. Also, preference must be given to an older request. With the help of timestamp, we select the oldest request. So, *PF* is inversely proportional to timestamp. Thus the terminology *PF* is defined as the ratio of the product of *BW* and *RRF* to t_s. It is expressed as $PF = (BW \times RRF)/t_s$.

2.2. Proposed Functional Model

The model proposed in this paper considers the situation described in Fig. 1. The sender S_i sets its semaphores to busy state just before sending the requests. Then it checks the receiver's status whether it is busy or idle. If the receiver is busy, the leader of S_i records t_{s_i} and corresponding RRF_i of the sender node and subsequently store them into a generic linked list.

The linked lists are represented with following fields – (a) for the leader of the receiver node: sender's id, the PF value and address to the next node, (b) for the leader of the sender node: sender's id, RRF of S_i , t_s value and address to the next node. The linked representations are shown in Fig. 2 for Fig. 1.

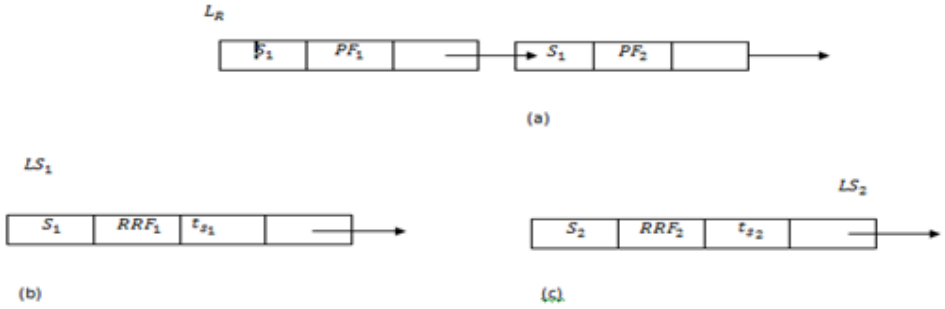


Fig. 2: Initial Linked Representation: (a) Linked List for L_R
(b) Linked List for LS_1 (c) Linked List for LS_2

the semaphore of R indicates idle, the L_R calculates the PF of the senders S_i that has been sending call requests and stores them into its own linked list. The receiver node then selects the sender that has maximum PF from the linked list and grants its requests. After the call has been granted, the t_{s_i} and the RRF_i of S_i are set again to initial values. This goes on dynamically for every node and the procedure is described by the following algorithm.

2.3. Algorithm:

Input: BW_i, initial values: t_{s_i} =0, RRF_i=0, flag = FALSE.
Output: $\max\,[PF(S_i)]$

Declaration: R=Receiver; S_i= i^{th} Sender ; $Avail_i$= Semaphore of S_i; $Avail_R$ =Semaphore of R; TS_{tot}= total timestamp; TS_{new}= new Timestamp; TS_{pre}= previous timestamp; L_R=Leader of the receiver; data storage=generic linked list;

```
Concurrent_Call_Request()
{
    Call();
    Receive();
    RRF_i =0;  /*setting the values to zero after the call has been granted*/
}
Call()
  {
        Avail_i=busy; /*setting its own semaphore to "busy" state before trying to connect to the
                                                                                    receiver*/
        While(Avail_R==busy) /*while the semaphore of the 'R' is in "busy" state the leader of Si
                                                                    increements  the ts and RRF values accordingly*/
      {
        TS_tot = TS_pre + TS_new ;
        RRF_i++;
 Record these values into the respective leader's data storage;
      }
  }
Receive()
 {
  if(count(S_i)>1) /* for multiple number of senders*/
  {
    While (i<=n) /*receive () executes until the computation of all PF for 'n' senders*/
     {
        L_R calculates the respective PF of the S_i ;
        PF = (BW_i × RRF_i/TS_tot) ;
        Each PF is then stored into the data storage of L_R;
      }
    calculate max [PF(S_i)] from data storage;
    flag=TRUE;
  }
```

Time Complexity: The algorithm executes in $O(n)$ time for both functions call ( ) and receive ( ).

2.4. Example

Initially the random available BW for call handling by S_1 and S_2 are assumed as 12 mbps and 10 mbps respectively. Similarly, the initial timestamp values for these nodes are randomly considered as 10 and 20 respectively.

STEP 1: Leaders are recording the respective values into its data storage as shown in Fig. 3.

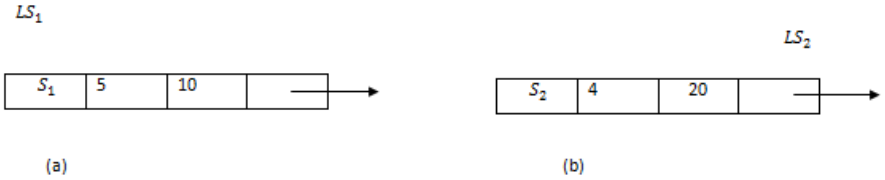


Fig. 3: After step 1: (a) Data Structure of LS_1 after entering the values; (b) Data structure of LS_2 after entering the values

STEP 2: L_R calculates PF of each S_i and stores them into its corresponding data storage as shown in Fig. 4.

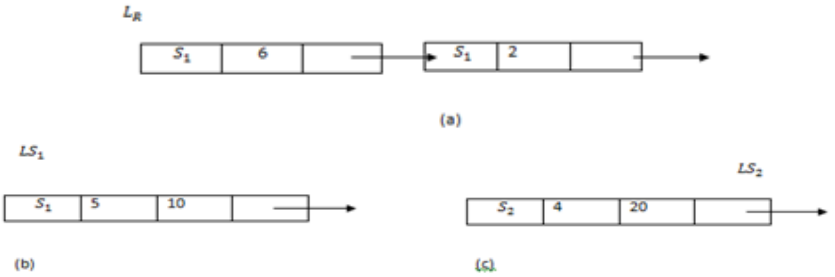


Fig. 4: After step 2: (a) Linked List for L_R (b) Linked List for LS_1 (c) Linked List for LS_2

STEP 3: Leader of Receiver L_R calculates the maximum of the PF from its data storage and then forwards the call to the respective receiver. Here, PF (S_1) > PF(S_2). Hence S_1's request is granted first.

STEP 4: Set the values of RRF_1 of S_1 to initial value as 0 and it is shown in Fig. 5.

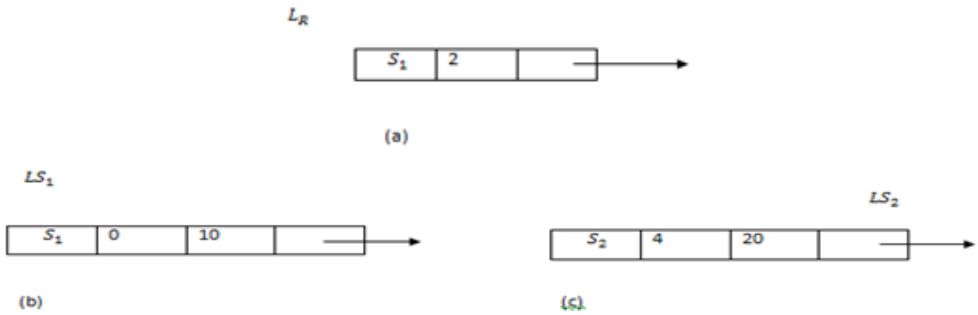


Fig. 5: After step 4: (a) Linked List for L_R (b) Linked List for LS_1 after its call has been accepted (c) Linked List for LS_2

2.5. Correctness

To prove the correctness of the proposed algorithm, a concept of progress graph [3] is used. The Progress graphs have intrinsic properties that are formalized by following postulates as given below.

P1: The concurrency state of a system defines a unique point in a progress graph.
P2: A transition from a state represented by a point p1 to a state represented by a point p2 is a ray rooted at p1 with direction p1$\vec{}$p2.
P3: A point is feasible if and only if it is not within a forbidden region. The forbidden region is the region that violates the constraints on the relative progress of the processes imposed by the signal events in progress graph.
P4: The time between two synchronizing events within each process is greater than zero.

Now, in our work, we summarize the events of the senders (S_1 and S_2) as follows in table 1:

TABLE 1: Events of the senders

S_1	S_2
P($Avail_r$)=waiting for $Avail_r$; V(S_i)=Signal S_i; V($Avail_r$)=Signal $Avail_r$; V(S_i)= Signal S_i ;	P($Avail_r$)=waiting for $Avail_r$; V(S_i)=Signal S_i; V($Avail_r$)=Signal $Avail_r$; V(S_i)= Signal S_i ;

Following these events for S_1 and S_2, the corresponding progress graph is shown in Fig. 6. Therefore, it is clearly seen that there is neither forbidden state, nor unsafe state as there is no starvation and deadlock. Thus the concurrency among the call requests is preserved.

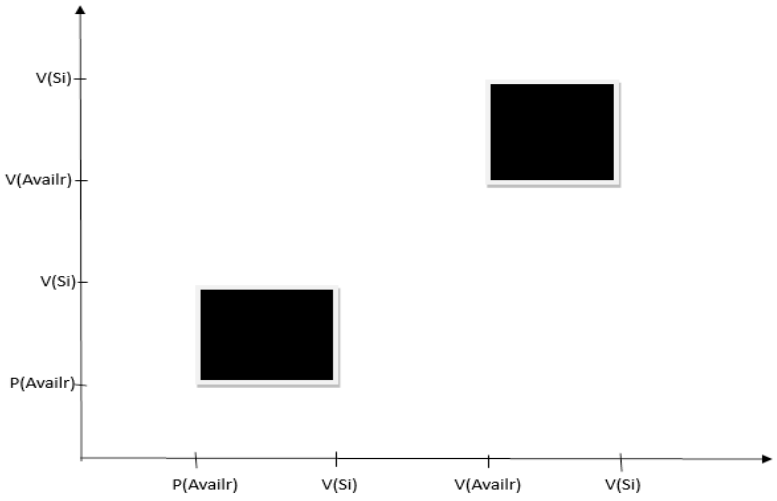


Fig. 6: Progress graphs

2.6. Discussion

Presently, while handling concurrent calls no preference is given to a sender node that is repeatedly trying to connect to a particular node. So, this may lead to indefinite waiting time for such node. For illustration, considering a sender node A sending a call request to a receiver node R repeatedly for (n-1) times and its request has not been processed yet. Now, sender node A sends the request for the nth time and concurrently a different node B sends a call request to same receiver node R for the first time. In such scenario, the system may usually process either of the requests without giving any preference to sender node A.

In our proposed approach, this indefinite waiting time can be handled by assigning a RPF factor to each node of the network. This RPF factor denotes the unsuccessful attempt of a sender node. Initially, it is set to 0 and incremented by 1 after each unsuccessful attempt. So, in the above scenario, RPF factor for A is n and RPF factor for B is 1.At the receiver node R , sender A is selected because of its greater RPF value

3. CONCLUSIONS

An efficient approach for concurrent call handling procedure is described in case of cluster based call scheduling for mobile networks. The semaphore concept is introduced here to resolve this concurrent issue. To determine the right sender among the alternatives, a priority factor (PF) is established. The several factors are determined to provide PF for a sender node. Furthermore, the proposed model executes in linear time. The performance of the model is analyzed with the help of progress graph. Moreover, we are extending our work with the help of logical clock to provide a compact and more efficient model for handling such concurrent events.

REFERENCES

[1] P.K.Guha Thakurta, Saikat Basu, Sayan Goswami and Subhansu Bandyopadhyay, “A New Approach on Cluster based Call Scheduling for Mobile Networks”, Journal of Advances in Information Technology, vol. 3, no. 3, August 2012.

[2] Silberschatz, Galvin and Gagne, “Operating System Concepts” , 7th Edition, February 8, 2005.

[3] Scott D. Carson and Paul F. Reynolds, Jr., “The Geometry of Semaphore Programs”, ACM Transactions on Programming Languages and Systems, Vol. 9, No. 1, January 1987, Pages 25-53.

[4] Zhengmin Kong, Liang Zhong, Guangxi Zhu, Li Yu, “A Novel Cluster-Based Routing Protocol and Cluster Reformation Criteria for Mobile Ad Hoc Networks”, 2010 International Conference on Computer Application and System Modeling (ICCASM 2010).

[5] P.K.Guha Thakurta, Rajarshi Poddar and Subhansu Bandyopadhyay, “A New Approach on Co-ordinate based Routing Protocol for Mobile Networks”, IEEE Advanced Computing Conference, February 2010.

AUTHORS

P. K. Guha Thakurta is an Assistant is an Assistant Professor at National Institute of Technology, Durgapur for the department Computer Science and Engineering with an experience greater than eight years. He published 5 International Journal and 11International Conference papers. His area of interest include DBMS, Network, Algorithm Analysis and Design, Formal language and automata

Jahnavi Katikitala is working as a software engineer at Samsung Research India, Bangalore. She has pursued her Bachelors of Technology from National Institute of Technology, urgapur. Her area of interest include Network, Algorithm Analysis and Design & Operating systems.

Misha Hungyo is pursuing her Masters in Computer Science and Engineering,in Motilal Nehru National Institute of Technology,Allahabad.She did her Bachelors from National Institute of Technology,Durgapur.Her area of interests are Computer Networking, Operating Systems, Data Structures and Algorithms.

Darakshan anwar is pursuing is working as a research engineer at C-DOT Bangalore.She has pursued her Bachelors of Technology degree from National Institute of Technology,Durgapur. Her area of interest include Network, Algorithm Analysis and Design.

MOBILE COMPUTING AND M-COMMERCE SECURITY ISSUES

Krishna Prakash[1] and Balachandra[2]

[1,2]Department of Information and Communication Technology, MIT Manipal
[1]kkp_prakash@yahoo.com,[2]bala_muniyal@yahoo.com

ABSTRACT

The radical evolution of computers and advancement of technology in the area of hardware (smaller size, weight, low power consumption and cost, high performance) and communications has introduced the notion of mobile computing. Mobile Commerce is an evolving area of e-commerce, where users can interact with service providers through a mobile and wireless network using mobile device for information retrieval and transaction processing. Mobile wireless market is increasing by leaps and bounds. The quality and speeds available in the mobile environment must match the fixed networks if the convergence of the mobile wireless and fixed communication network is to happen in the real sense. The challenge for mobile network lie in providing very large footprint of mobile services with high speed and security. Online transactions using mobile devices must ensure high security for user credentials and it should not be possible for misuse. The paper discusses issues related to M-Commerce security.

KEYWORDS

PKI, WPKI, Certificates, M-Commerce

1. INTRODUCTION

Mobile computing provides flexibility of computing environment over physical mobility. The user of a mobile computing environment will be able to access to data, information or other logical objects from any device in any network while on the move. To make the mobile computing environment ubiquitous, it is necessary that the communication bearer is spread over both wired and wireless media [2].

Mobile Computing technology evolved in various generation with changing technologies. The First generation (1G) mobile network was developed in USA and it was using Frequency Division Multiplexing technique (FDM). A data service was then added on the telephone network which was Cellular Digital Packet data (CDPD). The network could offer data rate of 19.2 kbps. The second generation (2G) mobile network is mainly Global System for Mobile Communication (GSM) and introduced in Europe and rest of the world. The network has dedicated data channels for data transmission.

The Third generation standards (3G) are developed by International Telecommunication Union (ITU) under International Mobile Telecommunication-2000 (IMT-2000) in order to create a global network. They are scheduled to operate in the frequency band around 2 GHz and offer data transmission rate up to 2Mbps. In Europe the ETSI (European Telecommunication Standard

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 205–216, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4920

Institute) has standardised UMTS (Universal Mobile Telecommunication System) as the 3GNetwork.

The ITU has stated the flow expected by 4G generation should be around 1GBPS static and 100 Mbps on mobility regardless of the technology or mechanism adopted.

The rapid development of mobile communication technologies and rapidly growing number of mobile devices result in fast growth of Mobile commerce.

1.1 Mobile System Infrastructure

One of the most widely deployed cellular infrastructures is GSM or 2G and its designers had several goals. Better quality for voice, higher speeds for data, international roaming, protection against charge fraud and eavesdropping. The UMTS or 3G promised advanced services such as mobile internet, multimedia messaging, video conferencing etc. UMTS standards were defined by an international consortium called 3GPP (Third generation partnership project) [3].

Fundamentals of a cellular system

The generic block diagram of a cellular system is shown in the Fig 1 below.

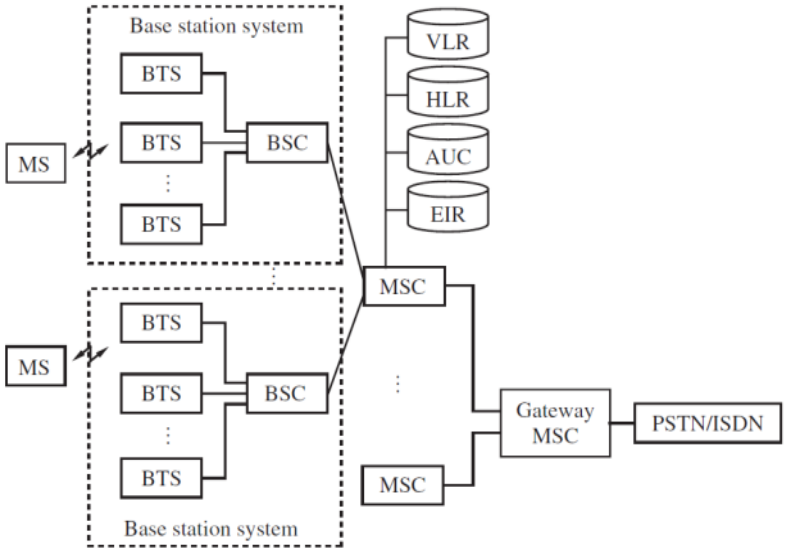


Fig 1: Cellular System

The basic geographical unit of a cellular system is called a cell is the geographical area covered by a transmitter. At the lowest level, a cell phone is connected to a base station (or base transceiver station) by a radio link. Multiple base stations are connected to and controlled by a base station controller. The connection between a base station and its controllers could be a microwave link, optical link in general any radio link. Multiple base station controllers and upstream are connected to Mobile switching centre. The Mobile Switching Centre (MSC) forwards an incoming call to the destination MSC. The MSC also keeps track of accounting and billing information. MSC are connected each other through wired networks such as Public Switched Telephone Network (PSTN).

The user has a subscription to some networks called as his home network. A one to one association between MSC and a network is maintained. An MSC has a database, called the Home Location Register (HLR) having information of all its subscribers. The data base contains the information of subscriber's mobile number, the services availed and a secret key stored in the mobile known only to the HLR. HLR also maintains the dynamic information of its roaming customers for charging. It includes the current location of a user and the cellular network used by the user [4].

A subscriber may avail the services of other networks (called as foreign networks) that have an agreement for roaming with subscriber's home network. Each cellular network also maintains a database called as Visitor Location register (VLR) of users currently visiting that network with the list of services the subscriber entitled to.2G technology introduced Subscriber Identity Module (SIM) card which stores three secrets used for cryptographic operations [5].

2. SECURITY IN POPULAR MOBILE NETWORKS

2.1 Security in GSM

There are two principal tasks involved for providing GSM Network security. They are:

a) Entity authentication and Key agreement
b) Message protection.

The integrity and encryption keys are agreed up on as a part of (a) and then they are used to protect messages between cell phone and base station.

a) Entity Authentication and Key Agreement

The GSM perform authentication to identify genuine users. The frequency of authentication is not specified, but the process is necessarily performed when the subscriber moves from one network to a new network. Fig 2 explains main steps involved in authentication.

1. Authorization request from Cell Phone: During authorization request step, the cell phone sends the encryption algorithm it can support to the base station and IMSI/TMSI number to the MSC. If the cell phone is away from its home network, the IMSI will be received by the MSC of the visited network. The latter communicates the IMSI to the MSC/HLR of the cell phones home network with a request to provide a challenge that will be used to authenticate by a cell phone.

2. Creation and transmission of authentication vectors:
   The IMSI obtained by the MSC is used to index the home location registers to obtain a shared key, Ki known only to the SIM and HLR of the home network. The MSC/HLR generates 128 bit random number, RAND, which functions as a challenge in the challenge-response authentication protocol. The two quantities XRES and Kc are computed as below.

XRES=A3 (RAND, Ki)
Kc=A8 (RAND, Ki)

Where, A3 and A8 are two keyed hash functions. XRES is the expected response in the challenge response authentication protocol. Kc is the encryption key. The HLR creates five authentication triplets, each seeded by freshly chosen random numbers. Each triplet is of the form-

<RAND, XRES, Kc>

The triplets are sent to the MSC of the home network by the HLR. If the cell phone is visiting a foreign network, the MSC forwards the triplets to the MSC of the visited network. Five triplets are sent so that four subsequent authentications may be performed without the need to repeatedly involve MSC/HLR of the home network.

The MSC sends the challenge (RAND) from the first triplet to the base station and it is forwarded to SIM on the cell phone.

3. Cell Phone response:

Once the SIM has received RAND, it computes SRES (Signed Response) similar to XRES. It can be computed by an entity with the knowledge of Ki, key shared between the SIM and HLR. The cell phone sends SRES to the base station and it is forwarded to MSC. The MSC compares if SRES is equal to XRES and if they are same MSC concludes that SIM knows Ki and identifies it as a genuine subscriber.

4. Computation/Receipt of encryption key:

The SIM computes Kc and MSC extracts Kc from its authentication triplet and communicates it to the base station. Further all communications between cell phone and base station are encrypted using Kc.

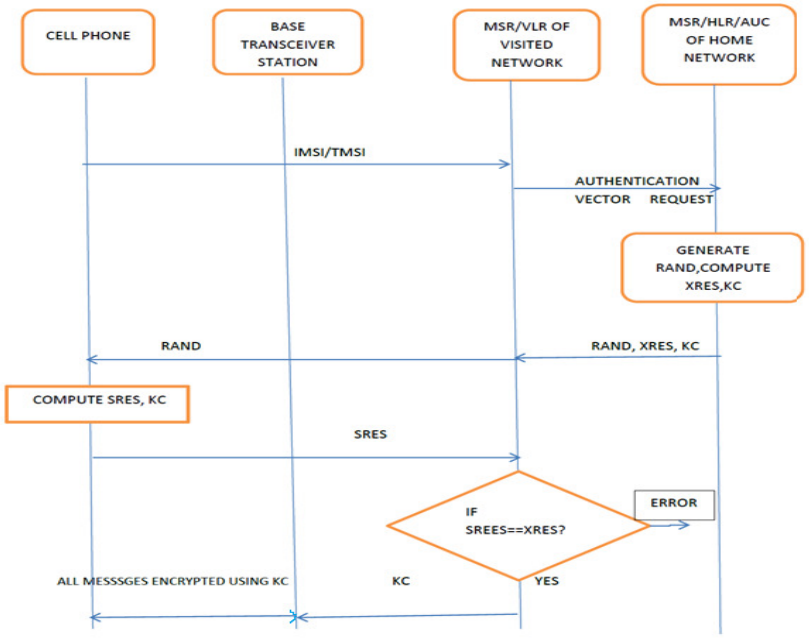


Fig 2: Authentication steps in GSM

b) Message Protection

Stream cipher technique is used to encrypt the message transmission between cell phone and base station. The key stream generator for this is denoted as A5. The key stream is a function of the 64 bit encryption key, Kc, and 22 bit frame number.

KEYSTREAM= A5 (Kc, FRAME_NUMBER)

For each frame transmitted, the frame number is incremented which changes the key stream for each frame sent during a call. Usually cipher text is generated by X-OR ing the plain text and the key stream.

Computation of the key stream and encryption do not require any static information stored in the SIM. Computation of XRES and Kc requires the subscriber authentication key, Ki. Hence the functions A3 and A8 must be supported by the SIM and A5 typically not.

2.2 Problems and drawbacks

There are some security shortcomings identified in GSM. The first flaw is related to authentication of the subscriber as illustrated in the following Fig 3 The system uses temporary identifier, Temporary Mobile Subscriber Identity (TMSI) to prevent the identity. If the VLR could not recognize or TMSI is lost, the IMSI is transmitted in plain text. There is no possibility of encrypting IMSI with A5, RAND is transmitted only after the successful authentication of the system is happened. This flaw may be exploited by using forged BTS and BSC. Unless the IMSI is transmitted in plain text subscriber is rejected. This type of attack is not common in principle in GSM networks and could be fought by a mutual subscriber-BSS authentication.

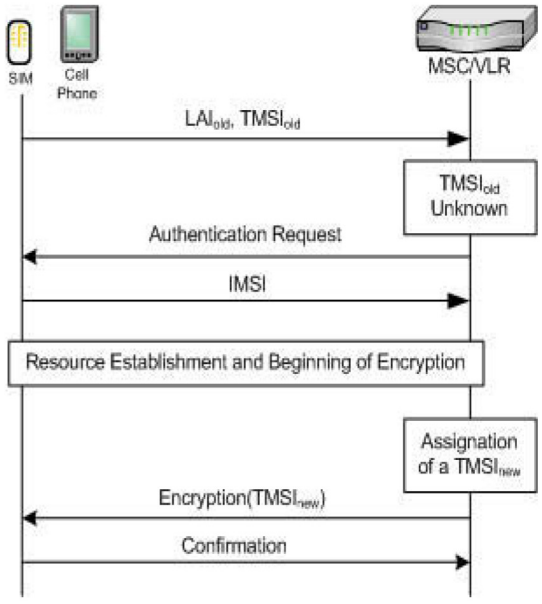


Fig 3: Unknown TMSI and plaintext IMSI transmission

In GSM, the SIM is authenticated to the network, but authentication of network is not carried out as a part of GSM protocol. This could result in false base station attack where an attacker poses as

base station by sending more powerful beacon signals than legitimate base station. The attacker may spoof the cipher mode command from base station and it suppresses the encryption in cell phone. As a result attacker may eaves drop entire communication.

The messages are encrypted only between the base station and cell phone, not beyond. The link between the base station and the base station controller is a microwave link and messages are transmitted in clear. Such links can be eaves dropped and defeating the purpose of GSM encryption.

Another flaw comes from SIM card cloning. If an attacker succeeds in cloning a SIM card and then turns a Mobile Network (MN) on, the network will detect two mobile devices with same identifiers at same time and will close the subscription and thus impeding identity thefts. Such attacks go undetected if the attacker is only interested about eavesdropping. If the intruder has access to secret key Ki and receives RAND may generate the encryption key Kc and passively decrypt the communication between cloned MN and attached BTS. This may be prevented by injecting copy protections and making them un clonable.

The major GSM security flaws find their origin in lack of any form of mutual authentication and plain text transmission of secrets. These flaws are identified and addressed in UMTS.

2.3 Security enhancements in UMTS

The 3G Security system define a higher security management for UMTS networks. New security provisions have been added such that detection of rogue base stations, network mutual authentication, strict control over the transmission of secret keys, longer encryption keys etc. GSM SIM card is replaced with more powerful chip called as USIM (Universal Subscriber Identity Module). Following features are built into UMTS to overcome the shortcomings of GSM.

1. False base station problem is impossible in UMTS, since each signalling message is individually authenticated and integrity protected.
2. GSM does not support mutual authentication of network and cell phone. In UMTS, as a part of mutual authentication protocol, the SIM card and the network agree on an encryption key and also a key for integrity protection of messages. To prevent replay attacks, the sequence numbers and nonce are used.
3. Data and signalling messages are encrypted. Both integrity protection and encryption are based on KASUMI-a 128 bit block cipher.
4. Messages on all wireless links are encrypted, not the link between cell phone and the base station. The algorithms for encryption and integrity can be negotiated between the SIM and the network.

2.4 Authentication and Key Agreement (AKA) in UMTS

The Fig 4 and Fig 5 discuss the AKA in UMTS by exploring the main difference with GSM.

a) Authorization request from cell phone:

This step is identical to that of GSM.

b) Creation and transmission of authentication vectors:

The HLR for the home network generates a random number, RAND functioning as a challenge in challenge-response protocol. Various keys such as “anonymity key (AK)”,

an integrity check key IK and a cipher key CK, a MAC and an authentication token (AUTN) are computed. The keys and expected response, XRES are derived using keyed hash functions F2, F3, F4, and F5 as follows.

XRES= F2 (RAND, Ki) …… ..(1)
CK= F3 (RAND, Ki)…… . . (2)
IK= F4 (RAND, Ki)……. (3)
AK= F5 (RAND, Ki)…… (4)

The HLR computes MAC (Message Authentication Code) using another keyed hash function F1.
MAC= F1 (RAND, Ki, AMF, SQN) ……. (5)

Here AMF is the Authentication Management Field containing the lifetime of the key. SQN is the secret sequence number known only to the HLR and SIM to maintain the synchronization between two. The HLR next creates an authentication token as follows.

AUTN= (SQN XOR AK, AMF, MAC)

Finally HLR produces five authentication vector quin tuplets as shown in Fig 2.6. each of the form,

(RAND, XRES, CK, IK, AUTN)

The SQN is incremented each time when a new authentication vector is created and RAND for each authentication vector is chosen a new.

The Authentication vectors are forwarded to the MSC/VLR of the visited network. Only once a single authentication vector is used for the authentication of SIM and MSC/VLR. The remaining authentication vectors may be used by MSC/VLR in future without the involvement of home network of the cell phone.

The RAND and AUTN of the first authentication vector is dispatched to the base station controller by MSC/VLR. The BSC forwards it to the SIM.

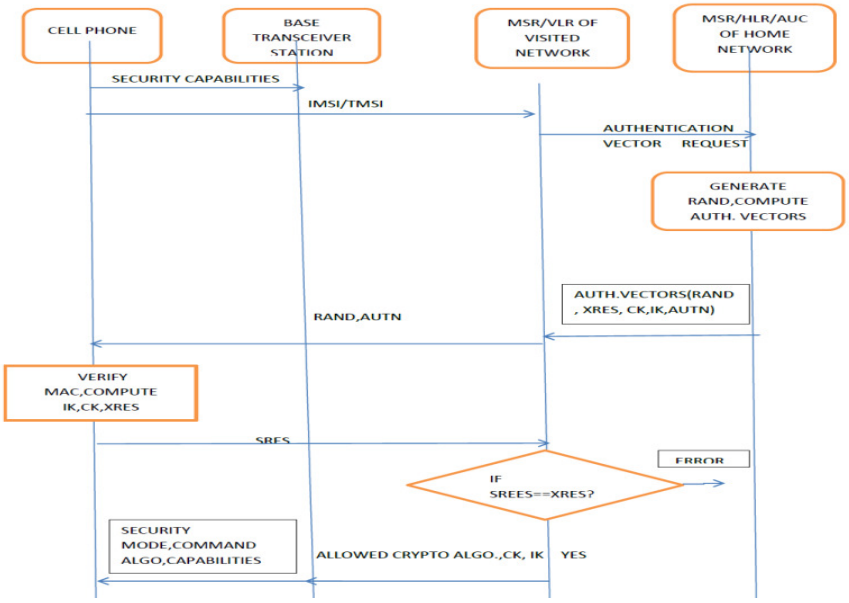


Fig 4: Authentication Protocol

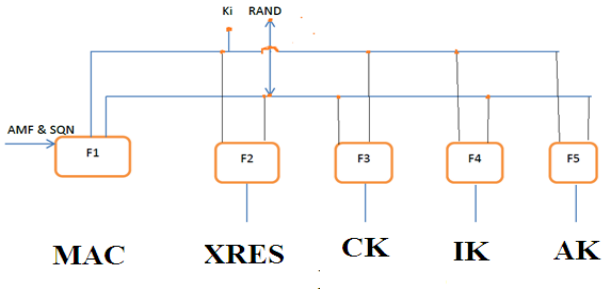


Fig 5: Authentication vector computation

c) Verification of authentication token and cell phone response

The SIM computes Authentication vector AK using equation (1) , the RAND it received and its copy of Ki. Also it retrieves first element of received AUTN, (SQN XOR AK) and it computes the value of SQN from (SQN XOR AK) XOR AK. After computation, it checks whether the difference between computed SQN and stored SQN is with in acceptable range. If it is OK, the SIM computes the MAC using equation (5). If the computed MAC and received MAC in AUTN matches , the SIM concludes that the authentication vector is created by HLR of the home network and authentication vector is fresh and not a replay. The SIM stores the SQN value it stored with the new value computed.

The SIM computes the response, SRES to the challenge, RAND (from HLR) using equation (1) and sends SRES to the MSC/VLR. The MSC/VLR compares both and if matches it proves that the SIM know Ki and completes authentication of SIM to the network.

At the SIM computes CK and IK and conveys these to the cell phone for providing encryption and integrity checking for all future communication between BSC and cell phone.

d) Agreement on Encryption and Integrity check Algorithms

The MSC/VLR sends a list of permissible MAC and encryption algorithms to the base station controller. The latter has received that from the first step and the BSC sends the list of supported algorithms back to cell phone. This message has an integrity check to prevent an attack from spoofed messages with weak options (may be no encryption). The BSC also receives CK and IK to be used for encryption and integrity protection of all messages between it and the cell phone.

3. MOBILE COMMERCE - RISKS, SECURITY AND PAYMENT METHODS

A Mobile Payment is defined as a payment for product or services between two parties for which a mobile device plays a key role in the realization of payment. In an M-Payment activity a mobile phone is used by the payer in one or more steps during banking or financial transactions. The ubiquity of cell phones together with the convenience it offers suggests that mobile payments will constitute an increasing proportion of electronic payments.

Mobile applications can be either be mobile web or native. Security issues in mobile web applications closely resemble those of traditional web applications because of homogeneity in underlying development technologies and protocols [6]

There are mainly two types of mobile payments as listed below [7].

1. Proximity Payment
2. Remote Payment

In Proximity Payment, the payer and payee are located nearby and they are very close to each other. Some examples for this category of payment are the customer paying the money using their plastic cards in a Point of Sale Terminal or Customers Cell phone making a payment in a vending machine.

In remote Payment, the payer and payee are located at different locations –for example they may be at different cities.

3.1 M-Payment Life cycle

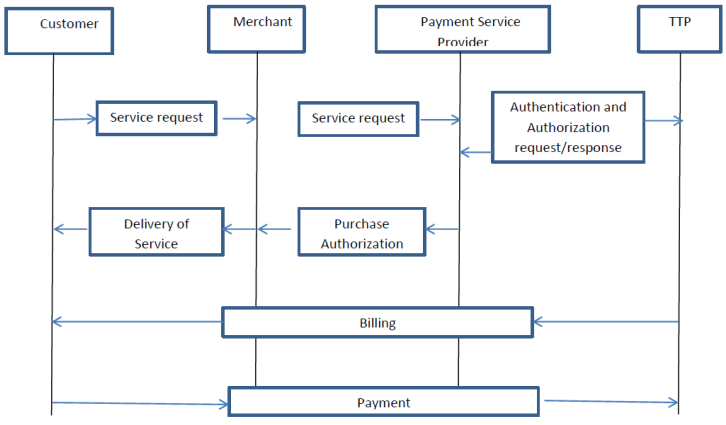


Fig 6: M-Payment life Cycle

Payment transaction in a mobile environment is very similar to a typical payment card transactions shown in Fig 6. It differs in the transport of payment detail involved i.e. wireless device using WAP/HTML based browser.

Mobile payment lifecycle has the following main steps.

1. Registration: Customer opens an account with payment service provider for payment service through a particular payment method.

2. Transaction: Transaction mainly comprised of following four important steps.

a) The desire of a customer is generated using a SMS or pressing a mobile phone button.
b) The content provider forwards the request to the payment service provider.
c) Payment service provider then requests a trusted third party to authenticate and authorize the customer.
d) Payment service provider informs content provider about the status of the authentication and authorization. If successful authentication of the customer is performed, content provider will deliver the requested goods.

3. Payment settlement: This operation can take place during real time, prepaid or post-paid mode. A real time payment involves the exchange of some form of electronic currency, for example payment settlement directly through a bank account. In prepaid type of settlement customers pay in advance using smart cards or electronic wallets. In post pay mode the payment service provider sends billing information to the trusted third party, which sends the bills to customers, receives money back, and then sends the revenue to payment service provider.

3.2 Wireless Public Key Infrastructure (WPKI) based M-Commerce Security System

Public key cryptography technique is used as backbone for the WPKI to provide security in m-commerce. The entire certificate management life cycle activities starting from certification creation, generation, storing, distribution and revocation of public key certificate is supported by an WPKI architecture. Fig 7 below illustrates various components existing in an integrated WPKI system [8]

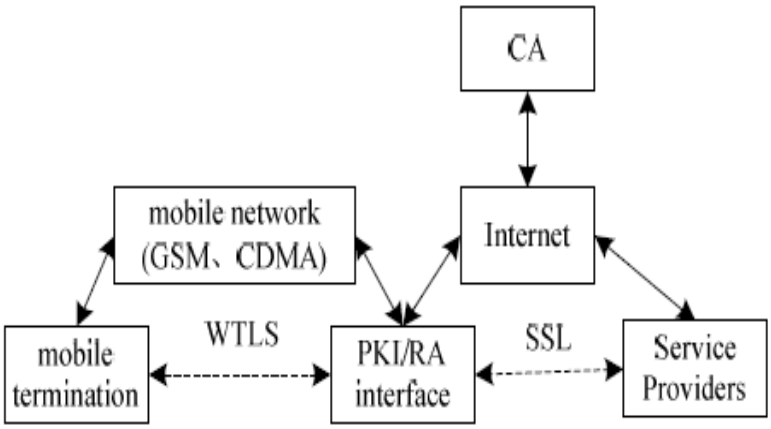


Fig 7: Components of M-Commerce security architecture

WAP is the key entity in an wireless environment for connecting the internet. WTLS is the lighter version of TLS and it is suitable for wireless environment. For the secure connection and communication between service providers SSL is used.

3.3 Secure Transmission Process between Mobile terminal and application server

The following Fig 8 depicts the architecture of secure transmission adopted in mobile commerce proposed by [9][10].

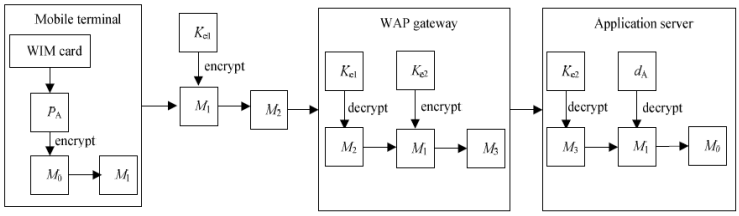


Fig 8: Secure data transfer between source and server

Following are the notations used in above diagram.

P_A- Public key of the application server
K_{e1}- The key between mobile terminal and WAP gateway
K_{e2}- The key between WAP gateway and application server
M_0- Message at mobile end
M_1- Message encrypted by P_A
M_2- Message encrypted with K_{e1}
M_3- Message encrypted with K_{e2}

The public key of application server P_A is used to encrypt the message M_0 and produces M_1. The message M_1 is then encrypted using the key K_{e1} of WAP gateway to get M_2 and sent to WAP gateway. The WAP gateway decrypts the message M_2 with Key K_{e1} to get original encrypted message M_1. The WAP gateway encrypts the message M_1 using K_{e2} and obtains M_3 and sends M_3 to application server. The application server will decrypt M_3 with K_{e2} and get M_1, and decrypts M_1 with its own private key and gets original message.

The above model has some draw backs.

1. The symmetric keys K_{e1} and K_{e2} needed to be generated and shared between WAP gateway and mobile device and Application server and WAP gateway securely.
2. The mechanism does not have any mechanism of verification of mobile user, WAP gateway and application server, so the chances of fraud is high. The following suggested modifications could be implemented for improved security.
3. Public key cryptography can be used to generate a pair of public key and private key and it can be used for both authentication and confidentiality.
4. The use of digital certificates can be used as a mean for authenticating mobile device and application server.

5. CONCLUSION

The widespread use of mobile devices now a day generates huge amount of revenues by reducing time and money needed for multiple purposes. The rapid development in mobile computing technology not only creates several opportunities for the business and also opens the door for doing disasters using misuse of technology. The information residing in the mobiles and integrity of the information, security of the information during its journey over the air security of the information with in the wireless network has to be given much importance.

Because of Mobile Computing or Mobile networks, M-Commerce has become reality today. The support of large number of cellular network service providers with competing speed made user to use his mobile device as a transacting module rather than simply using it for making calls.

REFERENCES

[1] Mahmoud Elkhodr, Seyed Shahrestani and Kaled Kourouche," A Proposal to improve the security of mobile banking applications", IEEE International conference on ICT and Knowledge Engineering, 2012
[2] Ashok K Talukder and Roopa R Yavagal, "Mobile Computing", TaTa McGraw Hill Education, January 2005
[3] Hua Ye, "Design and Implementation of M-Commerce system applied to 3G Network platforms based on J2ME", IEEE International conference on Electrical and Control Engineering, 2010
[4] Dharma prakash agrawal and Qing An Zeng, "Introduction to Wireless and Mobile Systems", Third Edition, Cengage Learning USA
[5] Hakima Chaouchi and Maryline Laurent maknavicius, "Wireless and Mobile Network Security", Second Edition, Wiley Publishers
[6] Anurag Kumar jain and Devendra Shanbhaug, " Addressing Security and Privacy Risks Mobile applications", IEEE Computer society, 2012
[7] Bernaard menezes, " Network security and cryptography", CENGAGE Learning, econd edition
[8] Feng Tian et al., " Application and Research of Mobile E-commerce security based on WPKI", IEEE International Conference on Information Assurance and Security, 2009
[9] CUI Jian-qi and YAO Dan-li, " New secure mobile Electronic commerce solution based on WAP [J].Application Research of Computers Vol.24 No.9 2007(9)
[10] ArunKumar Gangula et al., "Survey on Mobile Computing Security", IEEE Computer Society, 2013

INTELLIGENT ADAPTIVE LEARNING IN A CHANGING ENVIRONMENT

Guillaume Valentis[1] and Quentin Berthelot[2]

[1,2]Embedded Systems, ECE Paris Engineering School, France
valentis@ece.fr and berthelo@ece.fr

ABSTRACT

In order to develop ever more intelligent and autonomous systems, it is necessary to make them self-learning, since it is impossible to include in their program everything they may encounter during their life-cycle. In this research work, we aim at answering the following: if a system's environment is modified, how could the system respond to it quickly and appropriately enough? We achieve it by using reinforcement learning to allow the system to rate its decisions, then by developing adaptive learning algorithms for gain and loss rewards. The algorithms include probabilities' analysis providing to the system ability to adapt its knowledge through time and to respond to a changing environment. Simulations are made for a robot finding its exit in a labyrinth. Results show that reinforcement and adaptive learnings can have many useful applications by offering to a system a reliable possibility of evolution within complex environments in specific situations.

KEYWORDS

Reinforcement Learning, Neural Network, Autonomous Systems, Adaptive Learning, Changing Environment

1. INTRODUCTION

Present project concerns artificial intelligence in autonomous systems such as robots, and more precisely intelligence based on experience usually called reinforcement learning [1-5]. After development of high level mechanical capabilities for these systems with corresponding stability for trajectory following and robustness for handling adverse environment effects [6], embodiment of further cognitive capabilities became evident in a next step for higher efficiency with much larger adaptive response based on strong and redundant sensory background [7,8]. Within this extended setup, considerable attention is nowadays paid to giving autonomous systems access to "decisional" level for mastering themselves their own action [9,10]. Expectably, the most available results are obtained by usually expensive and time consuming processes often disqualifying real time response [11]. So here attention is focused on research of much shorter access to decision based on immediately available and global observations. The approach followed here is, contrary to some approaches which somehow antagonize environment change, to take instead advantage of this change when time elapses, implying that the system must adapt its knowledge. So the system should be learning during its whole life cycle because of environment changes happening around it. On the other hand, for all man-made autonomous and intelligent systems, one still cannot predict all situations they will encounter during their life cycle. Therefore, one must grant them some freedom to make their own decisions, and one possibility is that they may learn based on their gain consecutive to each particular decision

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 217–225, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4921

instead of other possible ones [12]. Rating the decision taken with higher or lower reward will enable the system to conclude on the decision having been good, average or bad, and thus to change its experience-based knowledge according to the actual gain provided by the environment (either the environment or an operator – human or machine). Therefore, the system would learn from its actions and correct its decisions from its mistakes [3,4,13].

The autonomous system must perceive the environment around it in order to make a decision before taking any action [3]. To illustrate the problem, the example of a robot in a labyrinth is considered here, where the overall path pattern stays the same, but an intersection can become an exit or a dead-end and conversely as time evolves.

2. THE MODEL

A system can learn in an unknown environment, thanks to reinforcement learning. Indeed, this consists in learning the pattern from experience: it has to face situations with rewards provided by the environment [3,14]. Typically the basic conceptual representation of such a situation is a tree-like structure where the system has to choose a branch to follow at each branching point for continuing its evolution. In a totally unknown environment with branching, the probability of taking a path instead of another one at an intersection is the same the first time the system encounters it. Thus, for m possible paths there is a probability $1/m$ to take each of them (it is a random neutral situation). Let the system choose a path and predict a gain upon taking this path. Its reward is the difference between final real gains and expected one. Depending on this reward, the probability to take again the same path will increase or decrease [15]. Here comes the interesting part: indeed the system is left learning for n times, n representing the number of times it would have made a complete journey in the environment (from the start to one end). After these n times, if the reward is consequently modified, in usual approach the system would take time to change its knowledge (if it is able to do it). Therefore the first statement made here is that no path can achieve a probability value equal to 0, and the least possible probability is fixed at a threshold value $P_{th} = 0.05$. It means that the system is allowed to take this decision 0.05 of the occurrences, giving it the possibility to explore this path and see if something new happened (if the reward has changed) or not since the previous visit. Therefore the maximum probability becomes:

$$P_{max} = 1 - 0.05 * (m - 1) \quad (1)$$

Then it is assumed that when a positive reward occurs, the gain curve must grow faster than the current probability to take this path. Indeed, taking again the example of a labyrinth explored by a robot, the robot has a 0.05 rate to go left where there is a dead-end, and has a 0.95 probability to go straight where it finds an exit. At some point it is decided to exchange the exit and the dead-end. When the robot takes the left path (0.05) with this very low rate one can predict that situation is worse than the straight path where there exists a 0.95 rate to take it. But once the path has been travelled the robot arrives at an exit, and the reward is positive. Therefore the probability has to be adjusted quickly to this change, but after certain limit the adjustment must be slower in order to stabilize the system. Thus one gets the following gain algorithm:

$$P_i = P_i + (1 - (m - 1) * 0.05 - P_i) * G * \varepsilon \quad (2)$$

Where Pi represents the probability at the previous intersection to take this path, m the number of different possibilities, G represents the reward (here only a positive reward can occur) and ε a learning coefficient (the smaller is the coefficient, the finer the learning curve is). The response time depends on ε, which has to be between 0.95 and 0.05 otherwise the program fails. Indeed, if $\varepsilon > 0.95$ or $\varepsilon < 0.05$, the gain can be superior to the maximum or inferior to the minimum

probability and make the system fail. We also need to consider the maximum gain factor as follows:

$$\varepsilon < P_{\max} / G_{\max} \quad (3)$$

With these elements, probability curves can be calculated for various values of ε. For $n=10$ runs with a gain $G=1$ one gets:

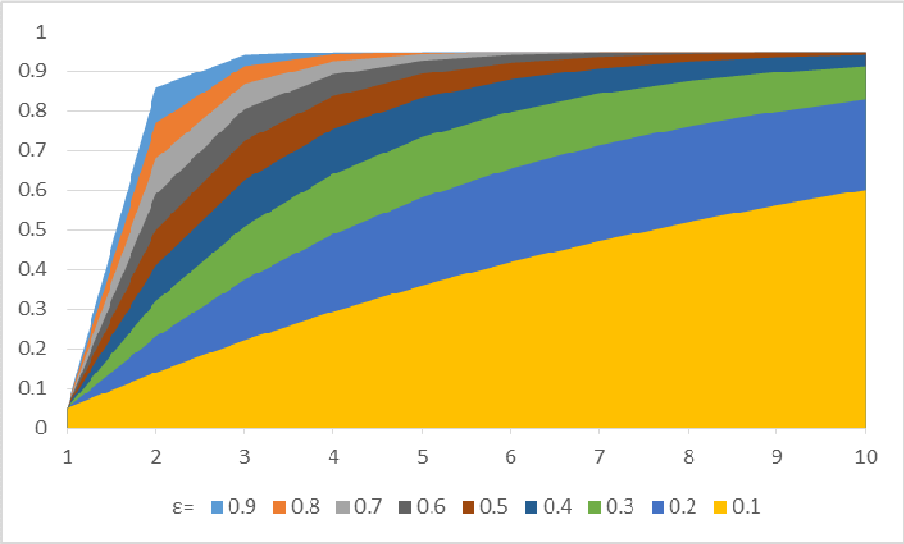


Figure1. Gain Probability Curves vs n for Different ε

On Figure 1, it is clearly seen that for $n=10$ the probability is around 0.6 for $\varepsilon=0.1$ and is already 0.9 for $\varepsilon=0.3$. This means it is a fast response for a radical change made on the environment (which happens when the exit and dead-end positions have been interchanged). Consequently this curve represents the gain curve only if the robot takes the path with the exit (previously a dead-end).

One should also provide a strong loss curve that may represent what is expected. When a loss occurs, the probability of uninteresting states is around 0.5, and one should make it change quickly and most smoothly around the extremities (0.95 and 0.05) as possible to have clear choices. Indeed, assuming that left path is a dead-end and straight path is an intersection, after n runs, the probabilities attained are respectively 0.05 and 0.95. If the guess of going straight is an exit and one gets an intersection, the reward is negative but actually it is not a so bad situation, therefore the probability must not change too much. Respectively the same problem occurs around 0.05 probability: if the loss was too high, the gain curve would not be strong enough to reverse the rate, even if this way was actually better.

The probability curve given in Figure 2 is split into two parts: the upper part (≥ 0.5) and the lower part (<0.5).

For $Pi \geq 0.5$:

$$P_i = P_i + (1-(m-1)*0.05-P_i-0.01)*G*\varepsilon \quad (4)$$

And for $Pi < 0.5$:

$$P_i = P_i + (P_i - 0.05) * G * \varepsilon \quad (5)$$

Where *Pi* represents the probability at previous intersection to take this path, *m* the number of different possibilities, *G* represents the reward (here only a negative reward can occur) and ε a learning coefficient (the smaller the coefficient is, the finer the learning curve becomes but the longer the training period is). Combining (4) and (5) in a digest algorithm with a programming test statement one gets:

$$P_i = P_i + ((P_i \geq 0.5)\,?(1 - 0.05*(m-1) - P_i - 0.01) : (P_i - 0.05)) * G * \varepsilon \quad (6)$$

Similar to gain probability curves, probability loss curves representation is shown on Figure 2 for various values of ε for $n=10$ runs with same gain $G=1$.

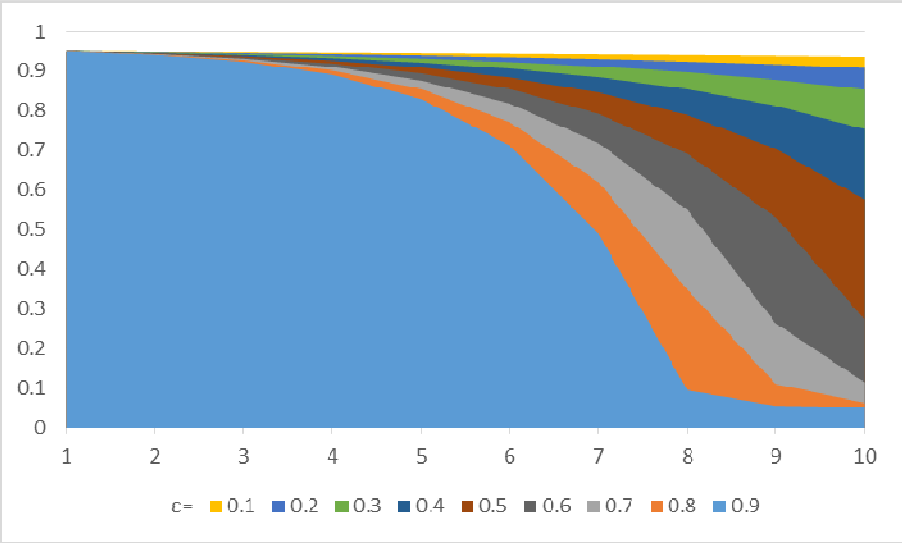


Figure 3. Loss Probability Curve vs *n* for Different ε

On Figure 2, it is clearly seen that for probability around *0.5* the changes occurs quickly while near *0.95* and *0.05* the curve decreases slowly. From both Figures 1 and 2, large (resp. small) probability values are reached faster with larger value of ε because slope is steeper due to higher convergence.

2. APPLICATION AND SIMULATION

Previous model will be applied to a labyrinth. Indeed this model is simple to understand, to study and to observe, but yet complex enough and on which one can easily make changes on the environment. All tests are performed on personalized software using algorithms explained in previous parts. The system first has to explore the labyrinth, the intersections when first met and get the random neutral probability for each possible way. It only knows the intersection at which it was located previously and its current position. The values of gain *G* used in simulation are:

- 0 for a dead-end;
- 1 for an intersection;
- 2 for an exit.

For the labyrinth the following pattern is used.

Table1. Symbols for the Different Labyrinth States

START EXIT WALL

And the system is represented by a small dark blue square. At each intersection the probability to take a specific way is recorded, and for all analysed examples the value is used to speed up the calculation.

As a first example, the basic two-way intersection is considered.

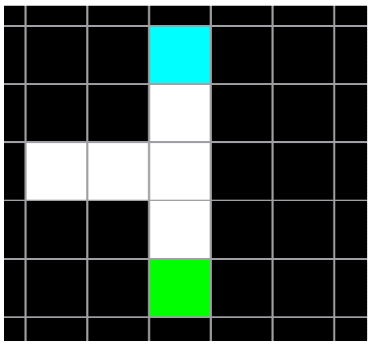


Figure3. Example 1 – Simple Intersection

As seen on Figure 2, there is a starting position, an intersection with two ways, on the left a dead-end and an exit standing forward. An exit is the highest gain while the dead-end is the lowest, so optimal observation is obtained with least runs.

Obtained probabilities at each sample are shown on Figure 4, on which it is observed that the system learns that it is clearly better to go forward at the intersection, therefore the probability to choose this way goes rapidly as high as *0.94*.

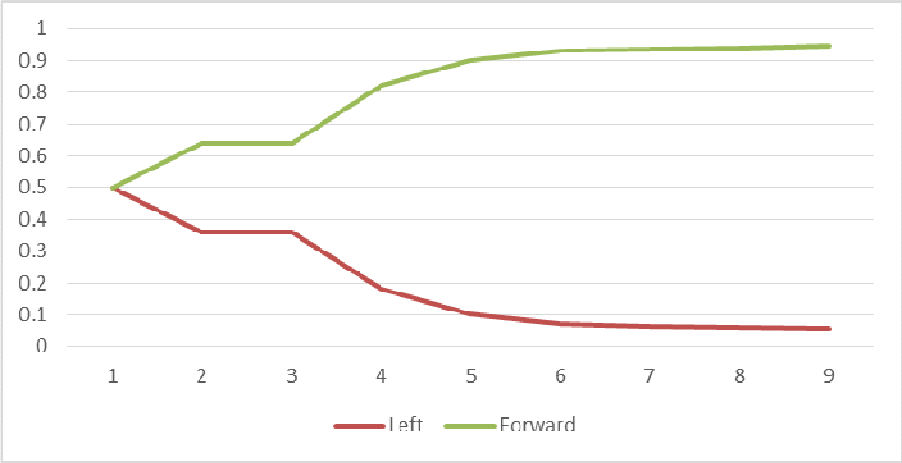


Figure4. Probability Curve until *0.95* Value for Taking the Exit (Forward Way) is Reached

Now the exit and the dead-end are exchanged while keeping the knowledge the system has already learned, see Figure 5.

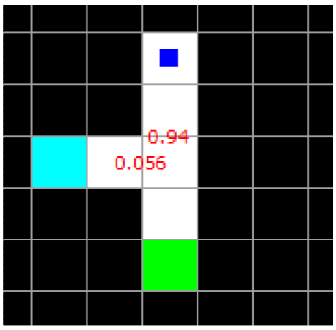


Figure5. Example 1 – Exit Switched with Dead-End

Evolution of probabilities in displayed on Figure 6.

Figure 6. Probability Curve vs Tries Number from Switched Initial Positions after Figure 5

It is seen that previously reached values are weakly modified up to the fourth try, where they drop down from *0.91* to *0.4* and raise steeply from *0.08* to *0.6*, and continue after a stop to increase further up to the same reversed values *0.94* and *0.06* they started from an initial state. Interestingly, the inversion is performed in the same number of tries (about 9) as the one the system had realized from its departure state. This indicates a strong robustness property of the proposed algorithm to the initial conditions, the convergence being here only dependent on parameter ε (at least for current value of ε).

For further testing analysis, the system maze is made more complex as shown on Figure 7. It still has a starting position and a two-way intersection (left-forward) but while the left leads to a dead-end again, the forward path leads to a two-way intersection (forward-right) with respectively an exit and a three-way intersection with, in successive order, an exit, a dead-end and a two-way intersection (right-forward) with an exit and a dead-end.

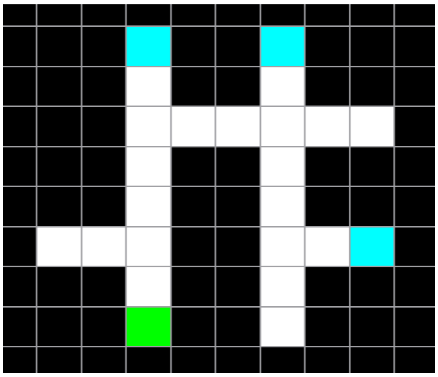


Figure7. New Testing Maze

System's probabilities when it browses the maze are displayed on Figure 8 vs run number.

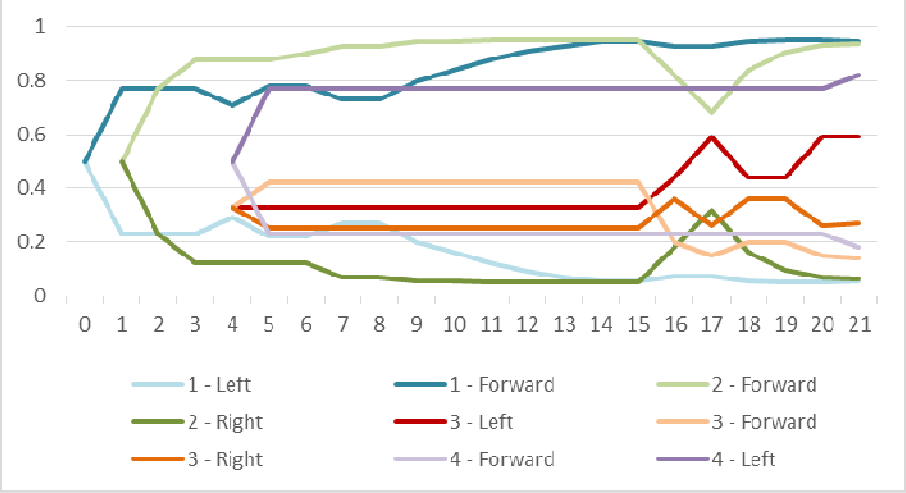


Figure 8. Probabilities Evolution vs try Number for the Different Paths

It is observed that all the different paths are not reached during the first run. This is logical since some paths are located after a possible dead-end or exit. Around the 15th run, abrupt changes are occurring on previously rather smooth curves. The probability limits are letting a chance to the system to look for other possible "good" paths. The curves are smoothing back to previous values as the results of exploration are impacting on system factor decision. On a general setting, many possible paths never reach their probability limits because the decision to take such path does not lead to the highest gain, and therefore it does not give a maximum chance for choosing it. The fact that some paths lead to other paths and so on in cascade, means that the probability tree will spread and give even less chance for each individual path. Final stabilized probabilities after the 21 runs are shown on Figure 9.

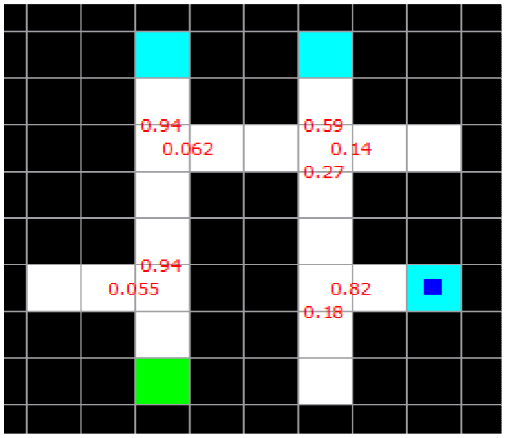


Figure 9. Stabilized Probability Distribution for the Various Paths after 21 Tries

From the probability values, clearly it is assumed that the system gets a better understanding of environment as it is nearby the start position, with a probability as high as *0.94* for the first gallery. This is to be compared with probability values of *0.59* for the first side exit and *0.82* for the second one once in the second gallery, which is explored with a probability of only *0.062*. So as it stands, the system over-weights neighbouring exits. Alternatively, observed dissymmetry is the mark that the system "remembers" its past in that it does not explore the same space to construct its trajectory from initial distribution. This implies the system has gained a further adaptation capability of the learning process for a better labyrinth structure understanding, and already stresses the efficiency of the simple proposed algorithm to rapidly orientate system exploration towards the most "efficient" exit.

4. CONCLUSION

In this paper, starting from reinforcement learning, we develop an adaptive learning method that provides to a system-robot the ability to evolve by its own and to take decisions depending both on its past experience and on changes occurring in its environment. The method is based on algorithms that use probabilities' analysis in order to incorporate gain and loss rewards.

According to this method, simulations are carried out to study the behaviour of the system-robot seeking the exit(s) of two types of maze: a simple one and a more complex one. The results of the simulations show that the system adapts smoothly and robustly when the positions of exit and dead-end are modified. It acquires a better understanding of the environment and its changes. Its possible behaviour when facing the different options, as interpreted by the probabilities, seems coherent with what could be expected from a human being.

Though simple, the tests here displayed of robot behaviour in a labyrinth are showing that reinforcement and adaptive learnings may have useful applications by giving to a system, with modest investment, reliable possibility of evolution within more complex environments in specific situations.

ACKNOWLEDGEMENTS

The authors express their gratitude to Dr F. Fauberteau for devoted coaching, to Pr. C. Duhart for guidance, to Pr. M. Cotsaftis for preparation of the manuscript, and finally to ECE Paris School of Engineering for having provided the setup in which the project has been developed.

REFERENCES

[1] L.P. Kaelbling, M.L. Littman, A.W. Moore, (1996) "Reinforcement Learning", a Survey, J. Artificial Intelligence Research, Vol.4, pp.237-285.

[2] S.F. Smith, C. Heng, H. Xi, (2003) "A Reinforcement Learning Approach to Production Planning in the Fabrication Fulfilment Manufacturing Process", Proc. Winter Simulation Conference, Vol.2, pp.1417-1423.

[3] T.C. Kietzmann, M. Riedmiller, (2009) "The Neuro Slot Car Racer: Reinforcement Learning in a Real World Setting", Proc. Intern. Conf. on Machine Learning and Applications (ICMLA '09), pp.311-316.

[4] Guang-Yi Cao, Bing-Qiang Huang, Min Guo, (2005) "Reinforcement Learning Neural Network to the Problem of Autonomous Mobile Robot Obstacle Avoidance", Proc. Intern. Conf. on Machine Learning and Cybernetics, Vol.1, pp.85-89, Guangzhou, China.

[5] P. Gerard, M. Butz, O. Sigaud, (2003) "Forward and Bidirectional Planning on Reinforcement Learning and Neural Networks in a Simulated Robot", Springer-Verlag, Ch.11, pp.179-200.

[6] Bin Yao, (1996) "Adaptive Robust Control of Nonlinear Systems with Application to Control of Mechanical Systems", PhD Thesis, Mechanical Engineering, UC Berkeley.

[7] I. Kanellakopoulos, (1993) "Passive Adaptive Control of Nonlinear Systems", Intern. J. Adaptive Control and Signal Processing, Vol.7, pp.339-352.

[8] S. Sastry, (1989) "Adaptive Control: Stability, Convergence and Robustness", Prentice Hall, Englewood Cliffs, NJ.

[9] M.R. Endsley, K.A. Ericcson, N. Charness, P.J. Feltovich, R.R. Hoffman (Eds.), (2006) "Expertise and Situation Awareness", The Cambridge Handbook of Expertise and Expert Performance, pp. 633–651, Cambridge Univ. Press.

[10] Haibo He, (2011) "Self-Adaptive Systems for Machine Intelligence", Wiley and Sons, New York.

[11] K. Kavi, R. Akl, A. Hurson, (2009) "Real-Time Systems: an Introduction and the State of the Art", Wiley Encyclopedia of Computer Science.

[13] K.O. Stanley, R. Miikkulainen, (2002) "Efficient Reinforcement Learning through Evolving Neural Network Topologies", Proc. Genetic and Evolutionary Computation Conference (GECCO-2002), pp.569-577, San Francisco.

[12] C. Gonzalez, V. Dutt, (2011) "Instance-Based Learning: Integrating Sampling and Repeated Decisions from Experience", Psychological Review, Vol.118 (4), 523-551.

[14] B. Bonet, H. Geffner, (2006) "Learning Depth-first Search: A Unified Approach to Heuristic Search in Deterministic and non-Deterministic Settings, and its Application to MPDS", Proc. 16th Int. Conf. on Automated Planning and Scheduling (ICAPS-06), pp.3-23.

[15] E. Mizutani, S.E Dreyfus, (1998) "Totally Model-free Reinfor-cement Learning by Actor-critic Elman Networks in non-Markovian Domains", Proc. IEEE World Congress on Computational Intelligence WCCI'98, pp.2016–2021, Alaska, USA.

AUTHORS

Guillaume Valentis is an engineering student at ECE Paris on his last year. He follows courses in embedded systems and robotics. He has studied at Concordia University and Dublin Business School.

Quentin Berthelot is an engineering student in ECE Paris, and is to earn his master's degree next year. He is studying embedded systems and robotics. He has attended classes at Concordia University and at UCI.

INTENTIONAL BLANK

A BELIEF REVISION SYSTEM FOR LOGIC PROGRAMS

Taher Ali[1], Ziad Najem[2], and Mohd Sapiyan[1]

[1]Department of Computer Science, Gulf University for Science and Technology, Kuwait
ali.t@gust.edu.kw, sapiyan.m@gust.edu.kw
[2]Department of Computer Science, Kuwait University, Kuwait
najem@cs.ku.edu.kw

ABSTRACT

Search is one of the most important needs of problem solvers. Usually the problem solvers suffer from retracing conclusions. If a problem solver cached its inference, then it would not need to retrace conclusions that it had already derived earlier in the search. By caching the inferences, the problem solver avoid throwing away useful results and avoid wasting effort rediscovering the same things over and over. In this paper we present a belief revision system for logic programs that can work under the non-monotonic logic.

KEYWORDS

Applications of justification-based truth maintenance systems, Belief revision systems, Truth maintenance systems, Justification-based truth maintenance systems, Incremental evaluation of tabled Prolog, Incremental tabulation for Prolog queries, Tabulation for logic programs, Memoing for logic programs.

1. INTRODUCTION

Truth Maintenance Systems [1], or Tms, are used within Problem Solving Systems [2], in conjunction with Inference Engines (IE) such as rule-based inference systems like Prolog (SWI-Prolog [3], Gnu-Prolog [4], B-Prolog [5], XSB [6], Ciao [7] and SICStus-Prolog [8]), to manage the inference engine's beliefs in given sentences as a Dependency Network. Figure 1 gives an an overview of Problem Solving Systems that uses Tms along with IE. A Tms is a knowledge representation method for representing both beliefs and their dependencies. A Tms is intended to satisfy a number of goals. One of these goals is the ability to remember derivations computed previously. It may happen that the same question is being asked from the problem solver over and over. If the previous knowledge is not cached when the question was answered for the first time, then the IE needs to re-compute the knowledge again and again. But if the previous knowledge was in the knowledge base, then there is no need for retracing the same knowledge. The use of Tms can avoid such retracing.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 227–231, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4922

Jtms is the simplest type of Tms where one can examine the consequences of the current set of assumptions. Jtms is a domain-independent belief revision system [9] which is usually coupled with an inference engine that does the actual inference work. Jtms operates on propositional objects and is used to record and maintain dependencies between deductive inferences. This can be done by representing deductive dependencies as a Jtms network.

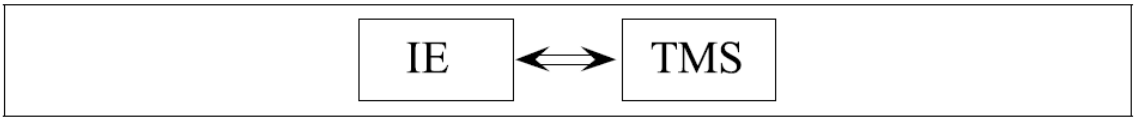


Figure 1: An overview of Problem Solving Systems

2. JTMS NODES AND JUSTIFICATIONS

The basic Jtms specification can be given in terms of two sets: the set of enabled assumptions (domain propositions) and the set of justifications. Propositions of the domain are mapped into nodes where each node is labeled either IN or OUT depending on whether or not it is currently assumed. Nodes corresponding to propositions that the system currently believes in are labeled IN while currently disbelieved propositions are labeled OUT. A node is labeled OUT by default but Jtms may label a node as IN in exactly two cases: either by a request from the inference engine or if there exists an active justification that supports the node.

In order to form a Jtms network the nodes are linked by justifications. A justification is a structure that is responsible for recording a single inference. A justification has two sets of nodes, the in -list and the out - list as its antecedent and a single node as its consequent. A justification is said to support its consequent node. Note that it is possible to have multiple justifications supporting the same node. An active justification is a justification where all the nodes in the in - list are labeled IN and all the nodes in out -list are labeled OUT. The consequent node of an active justification will be labeled IN while the consequent node of an inactive justification will be labeled OUT unless it has another active justification supporting it.

3. JTMS AND NON-MONOTONIC LOGIC

Search is one of the most important needs of problem solvers. Usually the problem solvers suffer from retracing conclusions. If a problem solver cached its inference, then it would not need to retrace conclusions that it had already derived earlier in the search. By caching the inferences, the problem solver avoid throwing away useful results and avoid wasting effort rediscovering the same things over and over. One of the Jtms goals, inherited from the Tms, is that Jtms is able to remember derivations computed previously. Jtms can do this by caching the inferences. This effort of Jtms can help in implementing incremental tabling[10] features for Prolog that will work with non-monotonic logic [11, 12] programs. The idea is that instead of remembering the end results as traditional memoing implementations does, the Prolog inference engine caches its inferences by the help of Jtms. By caching the inferences, Jtms will reflect any change in data through its network to keep the inferences updated. The responsibility of Jtms is answering queries correctly with respect to the contents of Jtms nodes and justifications at the moment the query is made.

Example 1

Figure 2 shows an example of an Sldnf-resolution [13] . The Sld-derivation [14] tree is for the query ?-bachelor(X) with respect to the Prolog program of Figure 2. There are two branches in the main proof structure with only one being successful. The only answer generated for the query ?-bachelor(X) is coming from the second branch. Figure 3 shows

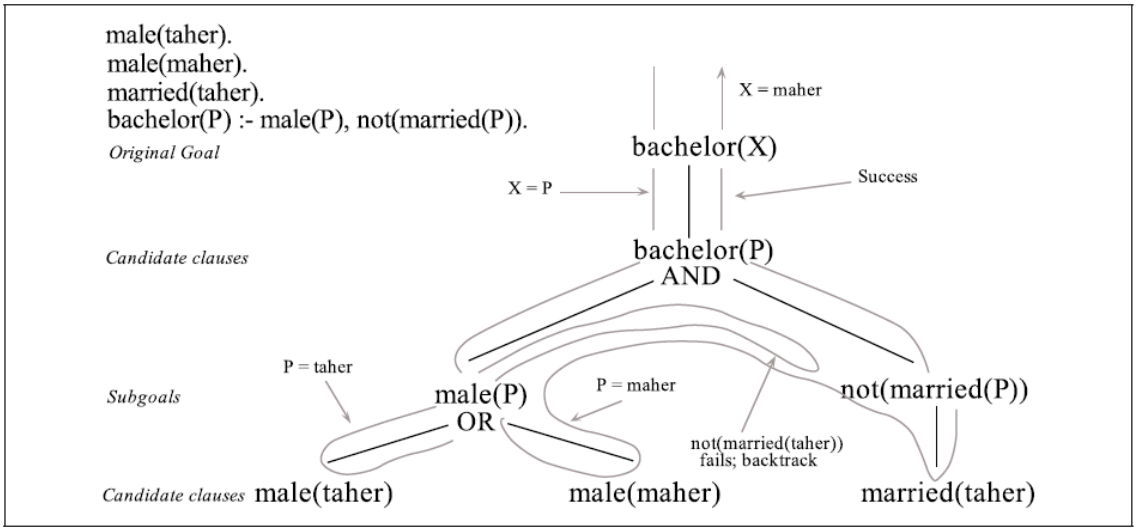


Figure 2: An example of Sldnf-dervation tree.

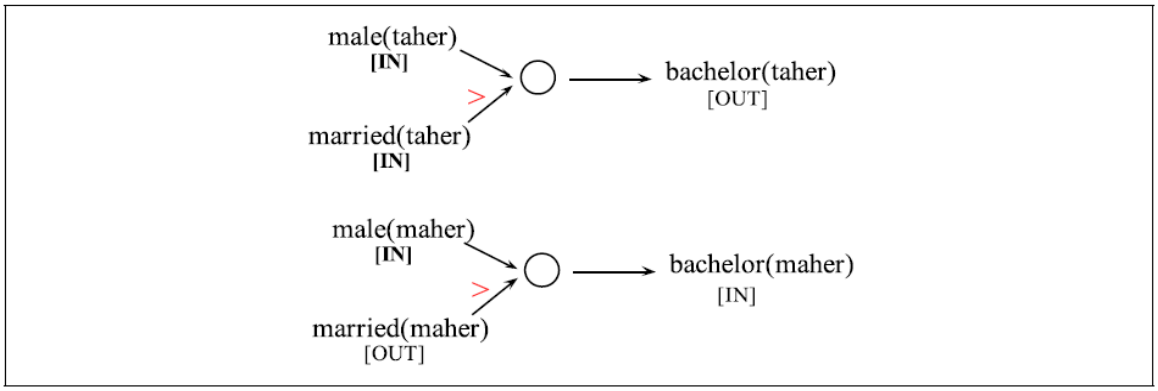


Figure 3: Jtms network for proof tree of Figure 2

an example of a justification network for the proof tree in Figure 2. Nodes are shown by printing their corresponding domain atoms. Justifications are shown as circles. The antecedents of a justification are identified by arrows pointing towards the justification while the consequent is pointing away from the justification. A negative literal in the antecedent (i.e. a member of the justifications out - list) is identified by placing a : sign on top of the arrow pointing to the justification. Figure 3 also shows the current labeling of the nodes. Labels that are printed in bold are specified by the inference engine while the

rest are assigned by the Jtms. Note that Jtms network of Figure 3 has two justifications that correspond to the two proof branches of Figure 2 proof tree, whether or not the proof branch was successful. This will allow the Jtms network to reflect any changes in data. (i.e. the query results are always correct and updated).

Coming back to the example of Figure 3, consider that married(maher) is asserted to the database of Prolog facts in Figure 2. Jtms reflect this change through it's network in order to keep the network updated. Jtms is capable of updating (revising) its belief bachelor(maher) without invoking the inference engine by propagating the changed value through the network. The resultant network is shown in Figure 4.

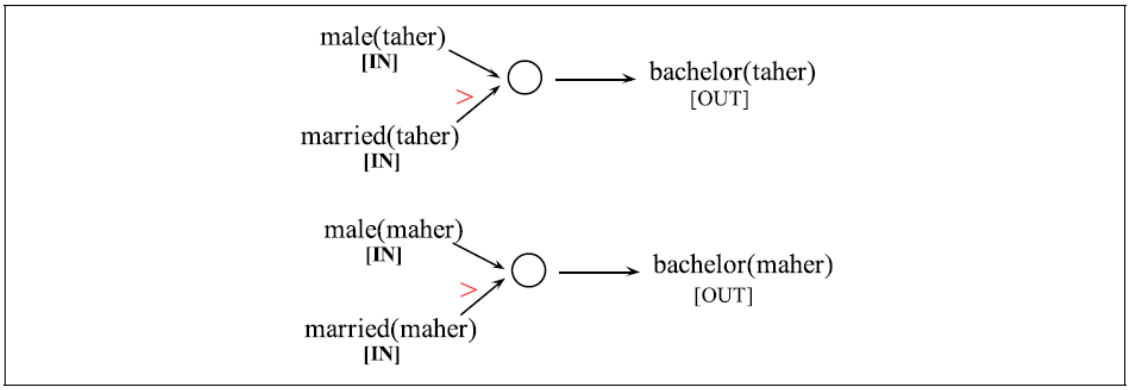


Figure 4: Jtms network of Figure 3 after asserting married(maher) to the database of Prolog facts in Figure 2.

4. CONCLUSION

The main idea of our approach presented in this papaer is to cache the proof generated by the deductive inference engine. The proof structure is converted into a justification based truth-maintenance (Jtms) network. Jtms saves the dependency between deduced facts and the facts used to make the deduction in order to be able to efficiently cache the proof structure. The system translates every successful branch of a query into a Jtms network that links the facts and the rule used in the branch to the answer generated by that branch. A justification is installed for each complete branch of the SLD-tree. When a query is re-evaluated, the system returns the answers of the query by collecting the IN consequences of each query's Jtms justification.When changes in database of facts take place, the system propagtes the effect of the changes through the Jtms network to ensure that the proof structure is both correct and complete.

REFERENCES

[1] Jon Doyle. A truth maintenance system. Artif. Intell., 12(3):231–272, 1979.
[2] Kenneth D. Forbus and Johan de Kleer. Building problem solvers. MIT Press, Cambridge,MA, USA, 1993.
[3] Jan Wielemaker, Tom Schrijvers, Markus Triska, and Torbjörn Lager. SWI-Prolog. Theoryand Practice of Logic Programming, 12(1-2):67–96, 2012.
[4] Daniel Diaz, Salvador Abreu, and Philippe Codognet. On the implementation of gnu prolog.TPLP, 12(1-2):253–282, 2012.

[5] Neng-fa Zhou. The language features and architecture of b-prolog. Theory Pract. Log. Program.,12(1-2):189–218, January 2012.
[6] Konstantinos F. Sagonas, Terrance Swift, and David Scott Warren. The xsb programming system. In Workshop on Programming with Logic Databases (Informal Proceedings), ILPS,page 164, 1993.
[7] Manuel V. Hermenegildo, Francisco Bueno, Manuel Carro, Pedro López-García, Edison Mera,José F. Morales, and German Puebla. An overview of ciao and its design philosophy. CoRR,abs/1102.5497, 2011.
[8] Mats Carlsson and Per Mildner. Sicstus prolog – the first 25 years. CoRR, abs/1011.5640,2010.
[9] Stuart C. Shapiro. Belief revision and truth maintenance systems: An overview and a proposal.Technical report, 1998.
[10] Diptikalyan Saha. Incremental evaluation of tabled logic programs. PhD thesis, Stony Brook,NY, USA, 2006. AAI3258884.
[11] Guido Boella and Leendert W. N. van der Torre. A non-monotonic logic for specifying and querying preferences. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, IJCAI, pages 1549–1550. Professional Book Center, 2005.
[12] Drew McDermott. Nonmonotonic logic ii: Nonmonotonic modal theories. J. ACM, 29(1):33–57, January 1982.
[13] David H. D. Warren. An abstract prolog instruction set. Technical Report 309, AI Center,SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, Oct 1983.
[14] Stefan Brass. Magic sets vs. sld-resolution. In Johann Eder and Leonid A. Kalinichenko, editors, ADBIS, Workshops in Computing, pages 185–203. Springer, 1995.

AUTHORS

Tahir M. Ali received his BSc and Ms from Kuwait University and PhD from University of Malaya. He is currently an Assistant Professor of Computer Science in G ulf University for Science and Technology, and also serving as the IT director. His main research interest is in field of Artificial Intelligence (AI), in particular, logic programming and scheduling algorithms.

Ziad H. Najem received his BSc from Kuwait University and Ms and PhD from University of Illinois at Urbana-Champaign. Prior to joining the Department of Computer Science at Kuwait University in 1999, Dr. Najem worked as a Scientific Resear cher at Kuwait Institute
for Scientific Research.

Mohd Sapiyan Baba is currently a Professor of Computer Science in Gulf University for Science and Technology, Kuwait. He was a lecturer in University of Malaya for more than 30 years, teaching Mathematics and Computer Science courses, and supervised numerous students for their research projects at undergraduate and postgraduate levels. His main research interest is in field of Artificial Intelligence (AI), in particular, the application of AI in Education

INTENTIONAL BLANK

DETECTION OF BLOOD VESSELS AND MEASUREMENT OF VESSEL WIDTH FOR DIABETIC RETINOPATHY

S.Sukanya, S.Abinaya and Dr.D.Tamilselvi

Department of CSE, Thiagarajar College of Engineering Madurai.
sabinayavasan@tce.edu, dtamilselvi@tce.edu

ABSTRACT

The proposed method measures the retinal blood vessel diameter to identify arteriolar narrowing, arteriovenous (AV) nicking, branching coefficients to detect early diabetic retinopathy. It utilizes the vessel centerline and edge information to measure the width for a vessel segment. From the input retinal image, the vascular network is extracted using the local entropy thresholding method. The vessel boundaries are extracted using sobel edge detection method. The skeletonization operation is applied to the vascular network and mapping the vessel boundaries and the skeleton image. The branching point detection method is then performed to localize all crossing locations. A rotational invariant mask to search the pixel pairs from the edge image, and calculate the shortest distance pair which provides the vessel width (or diameter) for that cross-section. Variation in the width measurement identifies the diabetic retinopathy.

INDEX TERMS

Computer Vision, Image Processing, Edge Detection, Mapping, Diabetic Retinopathy.

1. INTRODUCTION

Computer vision is the branch of Artificial Intelligence that focuses on, well, Image Processing. In 1966-1980s Computer Vision Engineers explicitly works for the shift towards geometry and increased mathematical rigor. In 1990- 2000s face and broader recognition, statistical analysis in vogue and video processing starts [1]. Diabetic Retinopathy is an ocular manifestation of diabetes, a systemic disease, which affects up to 50% of all patients who have diabetes for 10 years or more [6]. Diabetic macular changes in the form of yellowish spots and extravasations that permeated part or the whole thickness of the retina were observed for the first time by Eduard Jaeger in 1856. It was only in 1872 that Edward Nettleship published his seminal paper ”On edema or cystic disease of the retina” providing the first histo-pathologic al proof of ”cystoid degeneration of the macula” in patients with diabetes. In 1876, Wilhelm Manz described the proliferative changes occurring in diabetic retinopathy and the importance of fractional retinal detachments and vitreous hemorrhages [3]. In 1950s, a number of clinical trials have characterized the natural history of DR and the efficacy and safety of DR treatment strategies [2]. DR is an apparent breakdown of the blood-retinal barrier (Cunha Vaz 1976, Krupinet al. 1978,

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 233–246, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4923

Klemen et al. 1980) [4]. Normally the circulating blood is separated from the extravascular compartment of the retina by tight encircling functional complexes between contiguous endothelial cells in the case of the intra retinal vessels (inner blood-retinal barrier) and between cells of the retinal pigment epithelium (outer blood-retinal barrier) [5]. A number of multi-centered clinical trials during the last ten years have contributed substantially to the understanding of the natural history of diabetic retinopathy and have established the value of intensive glycemic control in reducing both the risk of onset and the progression of diabetic retinopathy [5].Despite this intimidating statistics, research indicates that at least 90% of these new cases could be reduced if there was proper and vigilant treatment and monitoring of the eyes[7]. Diabetic Retinopathy often has no early warning signs, which may cause vision loss more rapidly. Diabetic Retinopathy tends to appear and progress in stages beginning with Mild Non-Proliferative Diabetic Retinopathy, progressing to Moderate Non Proliferative Diabetic Retinopathy, further advancing to Severe Non-Proliferative Diabetic Retinopathy and without proper attention developing in to the most severe stage, Proliferative Diabetic Retinopathy . Mild Non-Proliferative Diabetic Retinopathy is characterized by the presence of dot and blot hemorrhages [11] and micro aneurysms[10] in the Retina during your eye examination[8][9]. In Moderate Non-Proliferative Diabetic Retinopathy some of the small blood vessels in the Retina may actually become blocked. The blockage of these tiny blood vessels causes a decrease in the supply of nutrients and oxygen to certain areas of the Retina[8]. Severe Non-Proliferative Diabetic Retinopathy is characterized by a significant number of small blood vessels in the Retina actually becoming blocked, which results in areas of the Retina being deprived of nourishment and oxygen. A lack of sufficient oxygen supply to the Retina results in a condition called Retinal Ischemia[12][8]. Proliferative Retinopathy is the most severe stage of Diabetic Retinopathy and carries a significant risk of vision loss. The Retina responds to a lack of oxygen, by attempting to compensate for the reduced circulation by growing new, but abnormal blood vessels-a process called neovascularization [13][8].

2. SYMPTOMS OF DIABETIC RETINOPATHY

A symptom is something the patient senses and describes, while a sign is something other people, such as the doctor notice. For example, rowsiness may be a symptom while dilated pupils may be a sign. Diabetic retinopathy typically has no symptoms during the early stages. Unfortunately, when symptoms become noticeable the condition is often at an advanced stage. Sometimes the only detectable symptom is a sudden and complete loss of vision. The only way patients with diabetes can protect themselves is attend every eye examination their doctor tells them to go to. Based on the research done by the Professors Tapp RJ, Shaw JE, Harper CA et al it is identified that patient can be prevented from the loss of vision if Diabetic Retinopathy is detected earlier. The blood vessel in the retina bulges called microneurysms causes the vessel width to vary (when the vessel bulges) compare to the other vessels or the other eye of the human. This symptom will be taken as earlier sign for the Diabetic Retinopathy and sometimes burst in the blood vessels causes tiny blood spots (hemorrhages) in the retina. Therefore the methods are proposed below to measure the blood vessel width more accurately to identify diabetic Retinopathy earlier.

3. PROPOSED ALGORITHM

The proposed blood vessels width measurement algorithm based on the vessel edge and centerline. The major advantage of our technique is that it is less sensitive to noise and works equally for the low contrast vessels (particularly for minor vessels). Another advantage of our technique is that it can calculate the vessel width even when it is one pixel wide. The proposed

algorithm is composed of four steps. Since blood vessels usually have lower reflectance compared with the background, we apply the matched filter to enhance blood vessels with the generation of a MFR image. Secondly, an entropy-based thresholding scheme can be used to distinguish between vessel segments and the background in the MFR image. A length filtering technique is used to remove mis-classified pixels. Vascular intersection detection is performed by a branch point detection method. Then the vessel width is measured by proposed width measurement technique.

A. Preprocessing

The color components are considered separately because green channel exhibits the best vessel/background contrast while the red and blue ones tend to be very noisy in case of RGB. The proposed method work on the inverted green channel images, where vessels appear brighter than the back-ground. The two dimensional matched filter kernel is designe d to convolve with the original image in order to enhance the blood vessels.

B. Matched Filter

In [15], This concept is used to detect piecewise linear segments of blood vessels in retinal images. Blood vessels usually have poor local contrast. The two dimensional matched filter kernel is designed to convolve with the original image in order to enhance the blood vessels. A prototype matched filter kernel is expressed as

$$f(x, y) = -\exp(-x^2), \text{ for } |y| \leq L/2 \qquad (1)$$

where L is the length of the segment for which the vessel is assumed to have a fixed orientation. Here the direction of the vessel is assumed to be aligned along the y-axis. Because a vessel may be oriented at any angles, the kernel needs to be rotated for all possible angles. A set of twelve 16x15 pixel kernels is applied by convolving to a fund us image and at each pixel only the maximum of their responses is retained. The operator generates a template of values that are then applied to groups of pixels in the image.

C. Local Entrophy Thresholding

MFR (Matched Filtering Retinal) image is processed by a proper thresholding scheme in order to extract the vessel segments from the background. An efficient entropy-based thresholding algorithm, which takes into account the spatial distribution of gray levels, is used because an image pixel intensities are not independent of each other. specifically , we implement a local entropy thresholding technique ,described in [16] which can well preserve the spatial structures in the binarized /thresholded image. A local entropy thresholding technique, described in which can well preserve the spatial structures in the binarized/thresholded image. Two images with identical histograms but different spatial distribution will result in different entropy(also different threshold values).The co-occurrence matrix of the image F is an $P \times Q$ dimensional matrix ,that gives an idea about the transition of intensities between adjacent pixels, indicating spatial structural information of an image. Depending upon the ways in which the gray level i follows gray level j, different definitions of co occurrence matrix are possible. The co-occurrence matrix asymmetric by considering the horizontally right and vertically lower transitions.

$$t_{ij} = \sum_{t=1}^{P} \sum_{tk=1}^{Q} \delta \; where \; \delta = 0 \; otherwise \qquad (2)$$

the probability of co-occurrence of gray levels i and j can therefore be written as

$$p_{ij} = \frac{t_{ij}}{\sum_i \sum_j t_{ij}} \qquad (3)$$

if s, $0 \leq s \leq L - 1$, is a threshold. Then s can partition the co-occurrence matrix into 4 quadrants, namely A, B, C, and D

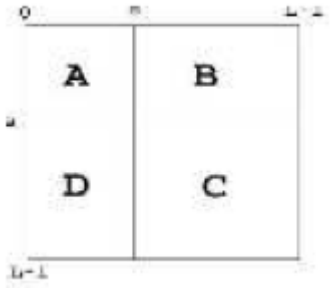


Fig.1.Quadrants of co-occurrence matrix [16].

Let us define the following quantities

$$P_A = \sum_{i=0}^{S} \sum_{j=0}^{S} p_{ij} \qquad (4)$$

$$P_C = \sum_{i=s+1}^{L-1} \sum_{j=s+1}^{L-1} p_{ij} \qquad (5)$$

Normalizing the probabilities within each individual quadrant, such that the sum of the probabilities of each quadrant equals one. The second-order entropy of the object can be defined as.

$$H_a^2(s) = -\frac{1}{2} \sum_{i=0}^{s} \sum_{j=0}^{s} p_{ij} \log_2 P_{ij} \qquad (6)$$

Similarly, the second-order entropy of the background can be written as

$$H_c^2(s) = -\frac{1}{2}\sum_{i=s+1}^{L-1}\sum_{j=s+1}^{L-1} p_{ij}\log_2 P_{ij} \qquad (7)$$

Hence, the total second-order local entropy of the object and the background can be written as

$$H_t^2(s) = H_a^2(s) + H_c^2(s) \qquad (8)$$

The gray level gives the optimal threshold for object back ground classification.

D. Vessel Edge Detection

The Sobel operator is used for edge detection algorithms. Technically, it is a discrete differentiation operator computing an approximation of the gradient of the image intensity function. At each point in the image, the result of the Sobel operator is either the corresponding gradient vector or the norm of this vector. The operator uses two 33 kernels which are convolved with the original image to calculate approximations of the derivatives - one for horizontal changes, and one for vertical.

E. Vessel Centerline Detection

Vessel skeleton is obtained by applying mathematical morphology reducing the vessel to a centerline of single pixel width. The perifoveolar capillaries were detected by the authors to locate candidate pixels(central part of vessel).Selection of vessel centerline candidates are done by using directional information provided from a set of four directional difference of Offset Gaussians filters. Connection of the candidate points are obtained in the previous step, by a region growing process guided by some image statistics. Validation of centerline segment candidates is based on the characteristics of the line segments; this operation is applied in each one of the four directions and finally combined, resulting in the map of the detected vessel centerlines.

F. Vessel Branch Point Detection

The vessel skeletons have to be converted into vessel segments separated by interruptions at the branching points. Segment start and end positions are determined as follows.Each of the centerline pixels on the vessel skeleton is analyzed within its 33 neighborhood, and branching points are detected as centerline pixels with more than 2 neighbors. The detection of vessel end points is required for the graph search and they are determined as the centerline pixels with only one neighbor.

G. Vessel Width Measurement

In [17][18],The vessel edge detected image and centerline detected image will be mapped to find the vessel width for a particular vessel centerline pixel position. For this purpose a pixel is selected from the vessel centerline image, considering the mask at its center. The mask is to find the potential edge pixels in any side of that centerline pixel position. Therefore, the mask is applied to the edge images only. Instead of searching for all the pixel positions inside the mask, width measurement method calculate the pixel position by shifting by one up to the size of the

mask and at the same time it rotate the position from 0 to 180 degrees. For increasing the rotation angle we use the step size (depending on the size of the mask) less than $\frac{180^{\circ}}{masklength}$. Hence, the method access every cell inthe mask using this angle.

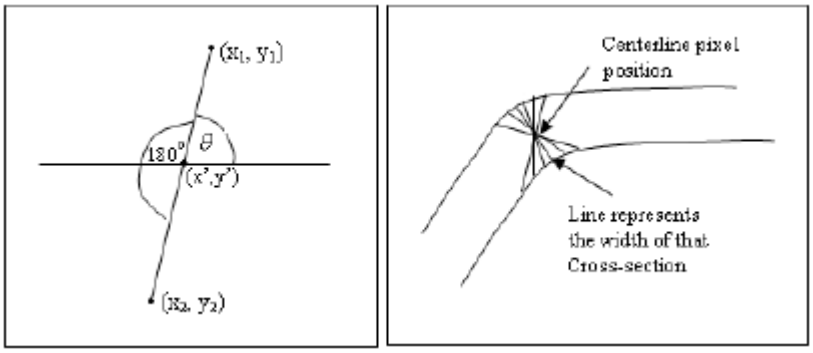


Fig. 2. Finding the mirror of an edge pixel(left) and width or minimum distance from potential pairs of pixels (right).

For each obtained position gray scale value is searched to check whether it is an edge pixel or not. Once we find an edge pixel then find its mirror by shifting the angle to 180 degree and increasing the distance from one to the maximum size of the mask (fig. 2) In this way the method produce a rotational invariant mask and pick all the potential pixel pairs to find the width or diameter of that cross sectional area.

$$x = x^{'} + r * \cos\theta \quad and \quad y = y^{'} + r * \sin\theta \qquad (9)$$

where (x′, y′) is the vessel centerline pixel position, r =1, 2, ..$\frac{marksize}{2}$ and θ = 0, .., 180. For any pixel position, if the gray scale value in the edge image is 255 (white or edge pixel) then we find the pixel (x2, y2) in the opposite edge(mirror of this pixel) considering θ = 0 + 180 and varying r this can be described in [17],[18]. After applying this operation we obtain the pairs of pixels which are on the opposite edges (at line end points) giving imaginary lines passing through the centerline pixels. From these pixels pairs, find the minimum Euclidian distance

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad (10)$$

the width of that cross-section. This enables us to measure the width for all vessels including the vessels one pixel wide (for which we have the edge and the centerline itself).

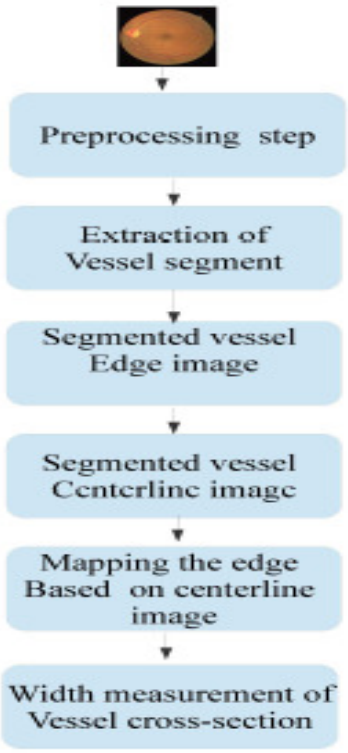


Fig. 3. The overall system for proposed vessel width measurement method

4. EXPERIMENTAL RESULTS

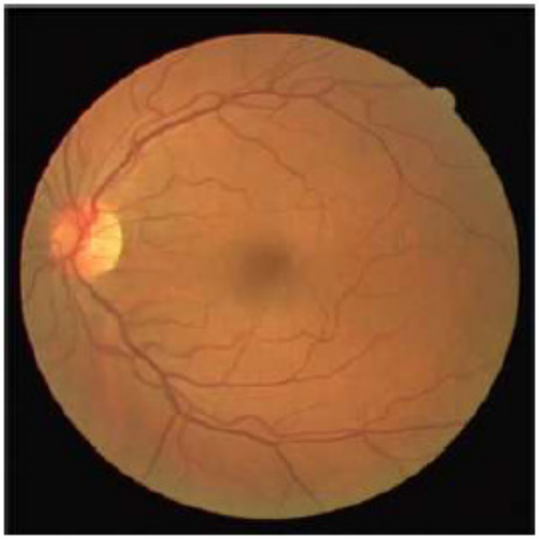


Fig.4. Original Fundus Image

Fig.3. represents the Fund us image is the inner lining of the eye made up of the Sensory Retina.

Fig.4. shows the Green Channel Conversion. The green channel conversion is helps to reduce the contrast of the image and used to separate the background and vessel segment.

Fig 5 shows the Gaussian filtering, to remove Gaussian noise and is a realistic model of defocused lens. Large values

Fig. 5. Green Channel conversion

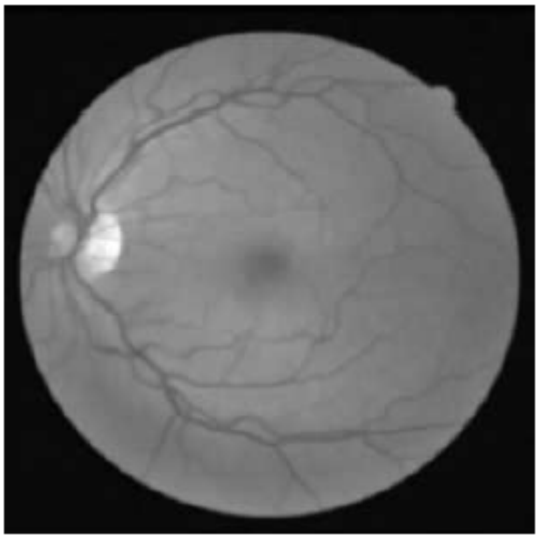


Fig. 6. Gaussian Filtering Image

for sigma will only give large blurring for larger template sizes. Noise can be added using the sliders. The radius slider is used to control how template size.

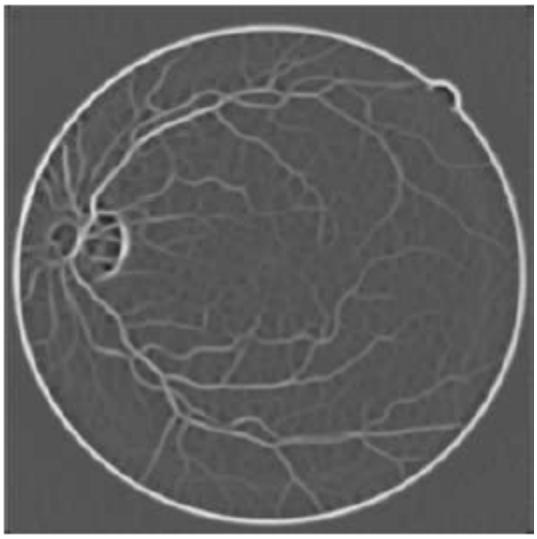


Fig. 7. Matched Filtering Image

The concept of matched filter detection is used to detect piecewise linear segments of blood vessels in retinal images. Blood vessels usually have poor local contrast as shown in Fig. 6. The two-dimensional matched filter kernel is designed to convolve with the original image in order to enhance the blood vessels.

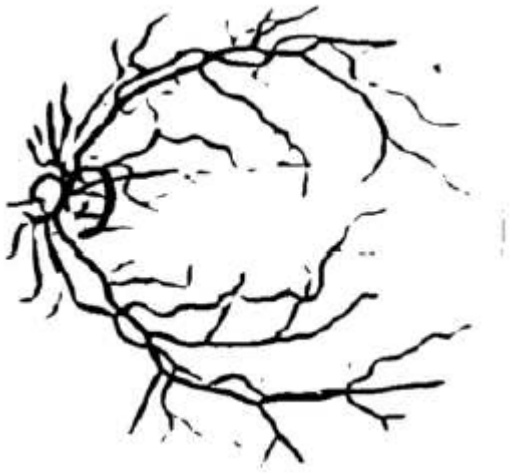


Fig. 8. Local Entropy Thresholding Image

Local entropy is a proper thresholding scheme in order to extract the vessel segments from the background as shown in Fig.7. It takes the spatial distribution of gray levels which is used to identify the pixel intensities. The pixel intensities are not independent of each other.

Fig.8. shows the length filtering which is used to produce a clean and complete vascular tree structure by removing misclassified pixels. It isolates the individual objects by using the eight-connected neighborhood and label propagation.

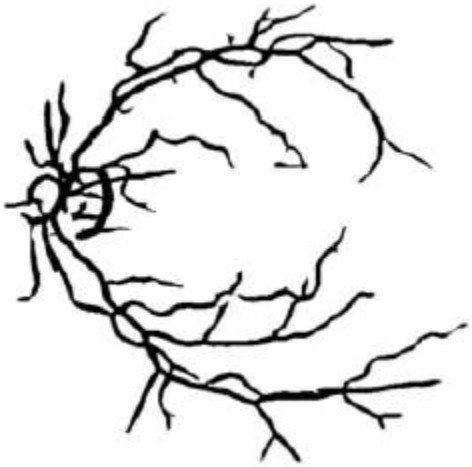


Fig. 9. Vascular Tree Structure

Fig.9. shows the edges of the detected vasculature tree structure (vessel boundaries).The edges of vessels are detectedusing the sobel edge detection method.

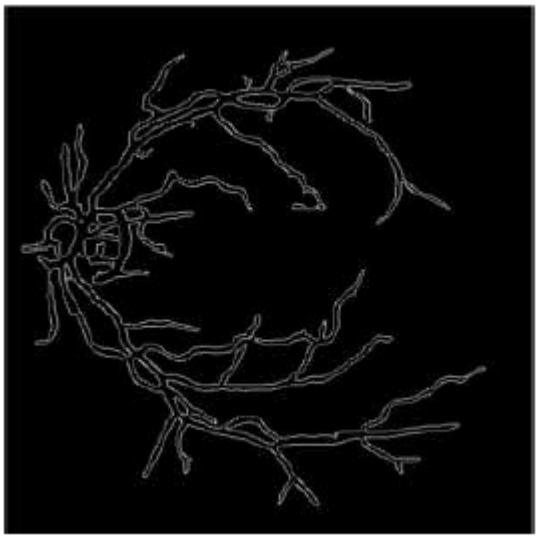


Fig. 10. Boundary of the vessel segment

The skeletonization of each vessel segment has shown in the Fig.10. It detects by using the morphological operators. This is also known as centerline detection of each vessel segment.

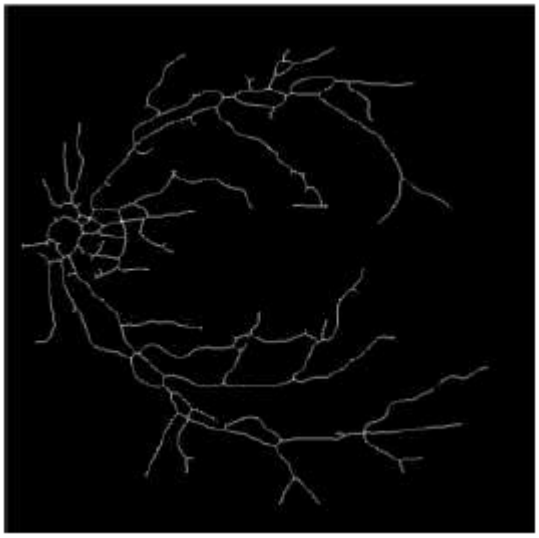


Fig. 11. Skeletonization of vessel segment (Centerline)

Fig.11. shows the mapping of vessel boundaries and centerline of retinal vessel. This is used to find the vessel width for a particular vessel centerline pixel position. The branching point of each vessel segements are detected and displayed in the Fig.12.The branching point are plotted in green color circle point. This points helps to determine the cross-section point of each vessel segment.

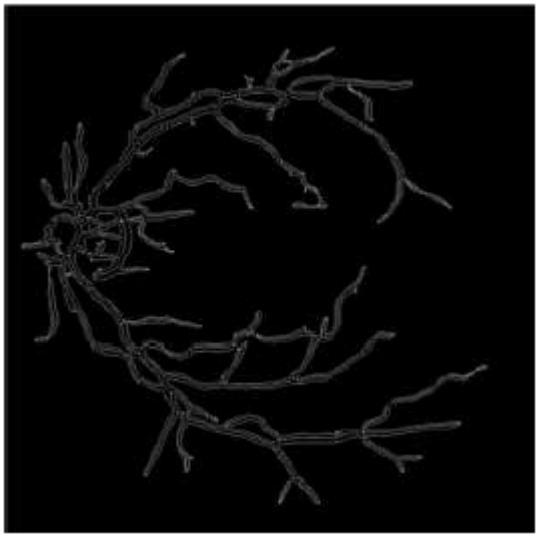


Fig. 12. Mapping of Boundary and skeletonization Image

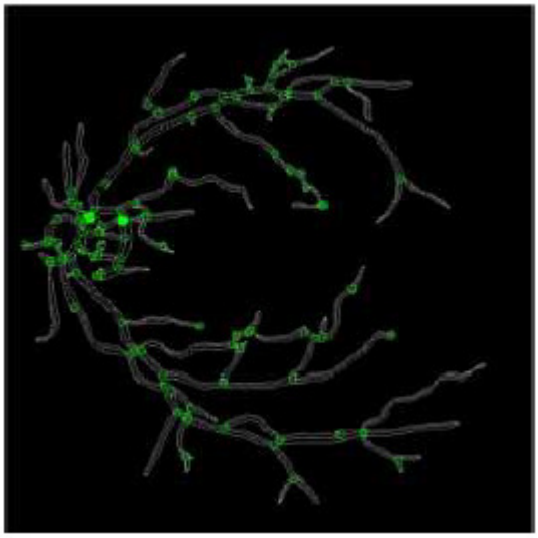


Fig. 13. Branching Point of each vessel segment

Fig.13. shows the branching points and ending points of each vessel segment. The branch points are plotted in blue color point and ending point of each vessel are plotted in red color point. And then the green color line segment shows the width or diameter of the particular vessel.

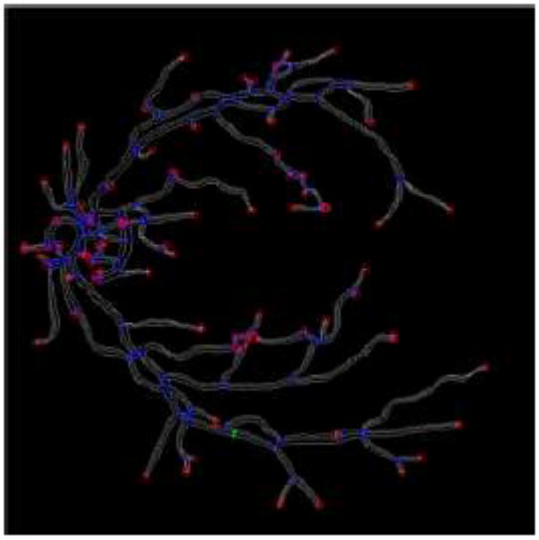


Fig. 14. Branch point, End point of vessel and Vessel width

Fig.14. shows the number of ending points and branching points which is plotted in the previous step. It also display the vessel width of particular vessel position.

```
No of ending points:82.
No of branches :104.
The point of vessel edge
  258.0000  260.0000
  480.0000  473.0000

The vessel width:
    7.2801

>>
```

Fig. 15. Vessel width

With the width calculated above, comparison based on vessel width calculation will takes place between manual width calculation and proposed method and the error(%) is displayed. The result is analyzed below based on the output obtained.

5. CONCLUSION AND FUTURE WORK

The proposed method assists for early detection and extraction of blood vessels and vessel width measurement in diabetic affected retinal image using the local entropy thresholding scheme and branch point detection. It reduces the computational simplicity compared to neural networks. It also identifies segmentation results for normal retinal images and images with obscure blood vessel appearance. In future classifying the artery and vein vessel segment and calculate the Arteriolar to Venular diameter Ratio (AVR) using this proposed vessel width estimation. A decreased ratio of the width of retinal arteries to veins [arteriolar-to-venular diameter ratio (AVR)], is well established as predictive of cerebral atrophy, stroke and other cardiovascular events in adults. Tortuous and dilated arteries and veins, as well as decreased AVR are also markers for plus disease in retinopathy of prematurity.

TABLE I
RESULT ANALYSIS

Cross-Section	manual width	proposed method	Error(%)
1	5.5369	5.3728	0.1641
2	5.5512	5.4721	0.0791
3	5.4381	5.3852	0.0529
4	5.219	5.1990	0.0200
5	5.1988	5.0990	0.0998
6	5.2733	5.1846	0.0887
7	5.3769	5.2980	0.0789
8	5.4316	5.3852	0.0464
9	5.45	5.3967	0.0533
10	5.4457	5.3798	0.0659

REFERENCES

[1] Derek Hoiem, David Forsyth, TA: Varsha Hedau Computer Vision University of Illinosis.

[2] Dibetic Retinopathy- http : //en.wikipedia.org/wiki/Diabetic retinopathy

[3] Wolfensberger TJ1, Hamilton AM.”DiabeticRetinopath–an Historical Review” Semin Ophthamill.2001 Mar; 16(1):2-7.

[4] Kalantzis G1, Angelou M, Poulakou-RebelakouE.”Diabetic retinopathy: an historical assessment”.Hormones (Athens). 2006 Jan-Mar;5(1):72-5.

[5] Alec Garner MD FRCpath-Department of Pathology, Institute of Ophthalmology,London EC] V 9A T ”Developments in the pathology of diabetic retinopathy: a review” in Journal of the Royal Society of Medicine Volwne 74 June 1981 427.

[6] Kertes PJ, Johnson TM, ed. (2007). Evidence Based Eye Care. Philadelphia, PA: Lippincott Williams & Wilkins. ISBN 0-7817-6964-7.

[7] Tapp RJ, Shaw JE, Harper CA et al. (June 2003). ”The prevalence of and factors associated with diabetic retinopathy in the Australian population”.Diabetes Care 26 (6): 17317. doi:10.2337/diacare.26.6.1731. PMID 12766102.

[8] T he Eye Center of Colorado -http://www.eyecarecolorado.com/diabeticretinopathy-denver.html

[9] A. Hoover, V. Kouznetsova, and M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, IEEE Transaction Medical Imaging, Mar.2000.

[10] Mahon, W. A., et al. ”Microaneurysms in Diabetic Retinopathy.” British Medical Journal (1971).

[11] Diabetic Retinopathy Vitrectomy Study Research Group. ”Early Vitrectomy for Severe Vitreous Hemorrhage in Diabetic Retinopathy: Four-Year Results of a Randomized Trial: Diabetic Retinopathy Study Report 5.”Archives of Ophthalmology 108.7 (1990): 958.

[12] Aiello, Lloyd Paul, et al. ”Vascular endothelial growth factor in ocular fluid of patients with diabetic retinopathy and other retinal disorders.”New England Journal of Medicine 331.22 (1994): 1480-1487.

[13] Schrder, S., W. Palinski, and G. W. Schmid-Schnbein. ”Activated monocytes and granulocytes, capillary nonperfusion, and neovascularization in diabetic retinopathy.” The American journal of pathology 139.1 (1991):81.

[14] Schrder, S., W. Palinski, and G. W. Schmid-Schnbein. ”Activated monocytes and granulocytes, capillary nonperfusion, and neovascularization in diabetic retinopathy.” The American journal of pathology 139.1 (1991): 81.

[15] Dr Caroline MacEwen. ”diabetic retinopathy”.Retrieved August 2, 2011.

[16] S. Chaudhuri, S. Chatterjee, N. Katz, M. elson, and M. Goldbaum,.Detection of blood vessels in retinal images using two dimensional matched filters,. IEEE Trans. Medical imaging, vol. 8, no. 3, September 1989.

[17] . R. Pal and S. K. Pal, .Entropic thresholding,. Signal processing,vol.16, pp. 97.108, 1989.

[18] U. T. V. Nguyen, A. Bhuiyan, L. A. F. Park, and K. Ramamohanarao, An effective retinal blood vessel segmentation method using multi-scale line detection, Pattern Recognit., vol. 46, no. 3, pp. 703715, 2012.

[19] U. T. V. Nguyen, A. Bhuiyan, L. A. F. Park, R. Kawasaki, T. Y.Wong, and K. Ramamohanarao, Automatic detection of retinal vascular landmark features for colour fund us image matching and patient longitudinal study, presented at the IEEE Int. Conf. Image Process., Melbourne, VIC, Australia, 2013.

A STUDY ON COMPUTATIONAL INTELLIGENCE TECHNIQUES TO DATA MINING

Prof. S. Selvi [1,] R.Priya[2], V.Anitha[3] and V. Divya Bharathi[4]

[1, 2, 3,4]Department of Computer Engineering,
Government college of Engineering, Bargur, India.
[1]sel_raj241@gmail.com,[2]priya231213@gmail.com,
[3]anithacs180294@gmail.com
[4]divyashree0129@gmail.com

ABSTRACT

Nowadays rate of growth of data from various applications of resources is increasing exponentially. The collections of different data sets are formulated into Big Data. The data sets are so complex and large in volume. It is very difficult to handle with the existing Database Management tools. Soft computing is an emerging technique in the field of study of computational intelligence. It includes Fuzzy Logic, Neural Networks, Genetic Algorithm, Machine Learning and Rough Set Theory etc. Rough set theory is a tool which is used to derive knowledge from imprecise, imperfect and incomplete data. This paper presents an evaluation of rough set theory applications to data mining techniques. Some of the rough set based systems developed for data mining such as Generalized Distribution Table and Rough Set System (GDT-RS), Rough Sets with Heuristics (RSH), Rough Sets and Boolean Reasoning (RSBR), Map Reduce technique and Dynamic Data Mining etc. are analyzed. Models proposed and techniques employed in the above methods by the researchers are discussed.

KEYWORDS

Data Mining, GDT-RS, RSH, RSBR, Map Reduce, Dynamic Data Mining, Rough Set Theory.

1. INTRODUCTION

It is a great challenge to deal Big Data. The data sets are characterized in terms of huge volume in quantity, high variety in type or classification, velocity in terms of real time requirements and constant changes in data structure or user interpretation. Basically it is very tedious task to understand the data. Hence big data reflects into revolutionary change in research methodology as well as tools to be employed in various applications. The conventional database management tools which are present now are not suitable to big data processing applications. The challenges include Data Analytics, Data Access, Capture, Curation, Sharing, Storage, Search, Transfer and Visualization etc. Therefore we require Computational Intelligence to solve real world problems. Some of the computational intelligence techniques are Evolutionary Computing, Swarm Intelligence, Fuzzy Logic, Neural Network, Machine Learning, Genetic Algorithm and Rough Set Theory etc.

David C. Wyld et al. (Eds) : COSIT, DMIN, SIGL, CYBI, NMCT, AIAPP - 2014
pp. 247–259, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4924

Computational Intelligence or Soft Computing techniques are exploiting the tolerance of imprecision, uncertainty, Partial truth information. Due to inconsistencies it is very difficult to mine knowledge. The Rough Set Theory is a mathematical model proposed by Pawlak [1], [2] which deal with vagueness to a great extent. A rapid growth of interest in rough set theory and its applications is being seen now. It is one of the first non-statistical methods of data analysis.

The basic concept of rough set theory is the approximation of spaces. The subset of objects defined by lower approximation is the objects that are definitely part of the interest subset and the subset defined by upper approximation are the objects that will possibly part of the interest subset. The subset defined by the lower and upper approximation [3] is known as Rough Set. Rough set theory has evolved into a valuable tool used for representation of vague knowledge, identification of patterns, knowledge analysis and minimal data set.

In modern decision support systems, data mining is the most prevalent and powerful tool used for extraction of useful, implicit and previously not known information from large data bases. Many of these data mining tasks search for the frequently occurring interesting patterns. This is done by using machine learning techniques. Data mining is part of Knowledge Discovery in Databases (KDD) [4]. Data mining is a specific step in KDD that involve the application of algorithms for extracting hidden patterns from data. It is used to find knowledge in the form of rules that characterizes the property of data or relationships, patterns etc. The data mining systems are mostly designed using traditional machine learning techniques. Rough set theory is powerful data mining tool which is implemented to reduce data sets, to find hidden patterns and to generate decision rules. The main advantage of using rough set theory is that it does not need any preliminary information about data. Recently, Rough set theory finds an important place among the researchers in intelligent information systems.

Some of the applications in which the rough set theory is efficiently employed are in the areas of medicine, social networking, aerospace engineering, market analysis etc. This paper presents the rough set theory, basic concepts of data mining, and the techniques of data mining in which rough set theory is used to improve the performance of data mining. Basics of data mining are discussed in section 2. The rough set theory is presented in section 3. The various data mining techniques based on rough set theory presented in various research articles are discussed in section 4. The analysis of the techniques is also consolidated in this section. Finally the survey is concluded in section 5.

2. DATA MINING

Data mining is a process of querying and extracting useful information [5], hidden patterns and unknown data in a database.

Main goals of data mining for many organizations include detecting patterns to improve marketing capabilities, future predictions etc. Decision making is difficult when the size of data base increases a large and obtained from many sources and domains. Thus, consideration is also to be given for the integrity of data.

Partitioning data into groups, associating rule to data and ordering data are the main tasks of data mining. With ubiquitous computing infrastructure, volume of data is also increasing to a larger extent. Hence, it is very difficult to have manual analysis of data. Data mining is specifically uses techniques for extracting features from such database for decision making.

2.1 The knowledge discovery process

Knowledge Discovery process through data mining is divided into four: Selection, Pre-processing, Data Mining and Interpretation [5].

Selection is a process of creating a target data set. It is not that the entire data base is to undergo the data mining process, because of the fact that the data represents a number of different aspects of the unrelated domain. Hence, the very purpose of data mining is to be clearly specified.

Pre-processing is nothing but processing or preparing the data set that could be used for analysis by the data mining software. This further involves activities that resolve undesirable data characteristics like missing data, irrelevant non-variant fields and outlying data points. This pre-processing step results in generating a number of subsets of original set. All the data are converted into a format acceptable for data mining software. The above process of collection and manipulation of data in data mining process is called collection and cleaning.

Data mining is the process involved in analyzing cleaned data by mining software to obtain significant results such as hidden trends and patterns.

Data ⇨ Information ⇨ Knowledge ⇨ Wisdom ⇨ Intelligence

Fig.1. Activities of KDD Process

Fig.1 shows the activities of KDD process. The rapid growth in IT is because of the KDD process. The purpose of Data mining and knowledge discovery is to develop methodologies and tools for automating the data analysis and thus creating useful information and knowledge. This helps in decision making faster.

2.1.1 Steps of Data Mining:

The steps of data mining are: organizing data, determining desired outcomes, selecting tools, mining the data and pruning the result. The results show that only the useful ones are further considered.

2.1.2 Data Mining Technologies and Techniques:

Data mining process shown in fig.2 is an integration of different technologies such as database management, data warehousing, machine learning and visualization etc. Some of the submitted methods are traditional and established.

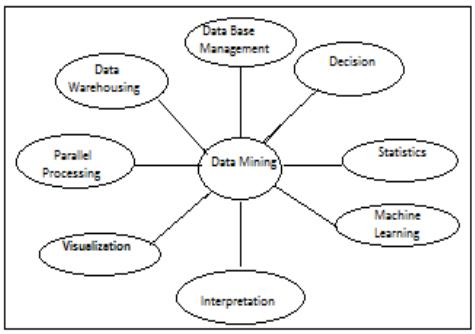


Fig.2. Data Mining Processes

2.1.3 Method Selection:

Some of the technical dimensions for the selection of data mining method, apart from the general considerations such as cost and support are given below:

- Uni-variate vs. multi-variate data.
- Numerical vs. categorical or mixed data.
- Explanation requirements or comprehensibility. Some tools give results, which are implicit to users (black box), while others can give causal and explicit representations.
- Fuzzy or precise patterns. There are methods such as decision trees, which only work with clear-cut definitions.
- Sample independence assumptions. Most methods assume independence of data patterns. If there are dependencies on the data patterns, it is necessary to remove or explore.
- Availability of prior knowledge. Some tools require prior knowledge, which might be not available. On the other hand, some others do not allow input of prior knowledge causing a waste of prior knowledge.

Other challenges come from lack of understanding of the domain problem and assumptions associated with individual techniques. Hence, data mining is not a single step. It requires multiple approaches to use some tools for the preparation of data.

3. ROUGH SET THEORY (RST)

Rough set theory, a new mathematical model developed by Pawlak in 1980s [1], [2], is an approach for imperfect or vague knowledge. This approach is used in the areas of knowledge acquisition, knowledge discovery, pattern recognition, machine learning and expert systems. Rough set theory provides means of identifying hidden patterns in data, finding minimal set of data, pointing out significant data, generating sets of decision rules from data. Rough set theory assumes that some form of information is associated with every object of universe. Objects are said to be indiscernible, if they are characterized by similar information. The mathematical basis of rough set theory is the indiscernibility relation exhibited by the above information.

The following are the listing of the concepts and research ideas of rough set theory:

Information System

Data for rough set based analysis are usually formatted into an *information system* IS= (U,A), where the set U is the universe of *objects* and the set A consists of *attributes*; any attribute a∈A is a mapping from U into a value set V_a. Subsets of U are *concepts*.

Indiscernibility Relation

It is the main concept in rough set theory, and is considered as a similarity relation between two objects or more, where all the values are identical in relation to a subset of considered attributes. Indiscernibility relation IND (A) is an equivalence relation [3].

Elementary Set

A set of indiscernible objects is known as elementary set. A union of elementary sets is referred to as crisp set otherwise the set is rough set

.

Lower and Upper Approximation

Vague concepts, in contrast to precise concepts, cannot be characterized in terms of information about their elements. Therefore, in the proposed approach, we assume that any vague concept is replaced by a pair of precise concepts, called the lower and the upper approximation of the vague concept [6].

Reducts

For a set B of attributes, one can look after an inclusion-minimal set $C \subseteq B$ with the property that IND(C) =IND (B), i.e., C is the minimal subset of attributes in B that provides the same classification of concepts as B. Such C is said to be a B-*reduct*.

Functional Dependence

For given A= (U,A), $C,D \subseteq A$, by $C \rightarrow D$ is denoted the *functional dependence* of D on C in A that holds iff $IND(C) \subseteq IND(D)$. In particular, any B-reduct C determines functionally D. Also dependencies to a degree are considered [2].

3.1 Algorithms used

To obtain the decision rules from the decision table, the algorithms LEM2 [7], [8], Explore [9] and MODLEM [8] are utilized. LEM2, Explore and MODLEM algorithms for rule induction are defined briefly as follows. These algorithms are strong for both complete and incomplete decision table induction.

3.1.1 LEM2 Algorithm

LERS [7] (LEarning from examples using Rough Set) is a rule induction algorithm that uses rough set theory to handle inconsistent data set, LERS computes the lower approximation and the upper approximation for each decision concept. LEM2 algorithm of LERS induces a set of certain rules from the lower approximation, and a set of possible rules from the upper approximation. The procedure for inducing the rules is the same in both cases [10]. This algorithm covers all examples from the given approximation using a minimal set of rules [11].

3.1.2 MODLEM Algorithm

Preliminary discretization of numerical attributes is not required by MODLEM. The algorithm MODLEM handles these attributes during rule induction, when elementary conditions of a rule are created. MODLEM algorithm has two version called MODLEM-Entropy and MODLEM – Laplace. In general, MODLEM algorithm is analogous to LEM2. MODLEM also uses rough set theory to handle inconsistent examples and computes a single local covering for each approximation of the concept [10]. The search space for MODLEM is bigger than the search space for original LEM2. Consequently, rule sets induced by MODLEM are much simpler and stronger.

3.1.3 Explore Algorithm

Explore is a procedure that extracts from data all decision rules that satisfy requirements such as strength, level of discrimination, length of rules and conditions on the syntax of rules. It may also be adapted to handle inconsistent examples either by using rough set approach or by tuning a proper value of the discrimination level. Induction of rules is performed by exploring the rule space imposing restrictions reflecting these requirements. The main part of the algorithm is based

on a breadth-first exploration which amounts to generating rules of increasing size, starting from one-condition rules. Exploration of a specific branch is stopped as soon as a rule satisfying the requirements is obtained or a stopping condition, reflecting the impossibility to fulfill the requirements, is met [11].

4. DATA MINING TECHNIQUES USING ROUGH SET THEORY

This section presents the various data mining techniques proposed by many researchers using the rough set theory.

4.1 Generalized distribution table and rough set system (GDT-RS)

GDT-RS is a soft hybrid induction system which helps to discover classification rules from databases with uncertain and incomplete data [12], [13]. The system is based on a hybridization of the Generalization Distribution Table (GDT) and the Rough Set (RS) methodology. The GDT-RS system can generate, from noisy and incomplete training data, a set of rules with the minimal (semi-minimal) description length, having large strength and covering all instances.

There are attributes, namely *condition* attributes and *decision* attributes (sometimes called class attributes) in a database. The condition attributes are used to describe possible instances in GDT, while the decision attributes correspond to concepts (classes) described in a rule. The GDT consists of three components: *possible instances*, *possible generalizations* of instances, and *probabilistic relationships* between possible instances and possible generalizations.

Possible instances are defined by all possible combinations of attribute values from a database. *Possible generalizations* of instances are all possible cases of generalization for all possible instances. The *probabilistic relationships* between possible instances and possible generalizations are defined by means of a probabilistic distribution describing the strength of the relationship between any possible instance and any possible generalization.

4.1.1 Simplification of the Decision Table by GDT-RS:

The process of rule discovery consists of the decision table preprocessing, including selection and extraction of the relevant attributes (features), and the appropriate decision rule generation. The relevant decision rules can be induced from the minimal rules (i.e. with the minimal length of their left-hand sides with respect to the discernibility between decisions) by tuning them (e.g. dropping some conditions to obtain more general rules which are better predisposed to classify new objects even if they do not classify properly some objects from the training set). The relevant rules can be induced from the set of all minimal rules, or from its subset covering the set of objects of a given decision table [14], [15]. A representative approach to the problem of generation of the so called local relative reducts of condition attributes is the one to represent knowledge to be preserved about the discernibility between objects by means of the discernibility functions.

It is obvious that by using the GDT one instance can be matched by several possible generalizations, and several instances can be generalized into one possible generalization. Simplifying a decision table by means of the GDT-RS system leads to a minimal (or sub-minimal) set of generalizations covering all instances. The main goal is to find a relevant (i.e. minimal or semi-minimal with respect to the description size) covering of instances still allowing us to resolve conflicts between different decision rules recognizing new objects. The first step in the GDT-RS system for decision rule generation is based on computing local relative reducts of condition attributes by means of the discernibility matrix method.

Relevant attributes are searched using bottom-up method instead of searching for dispensible attributes. Any generalization matching instances with different decisions should be checked by means of noise rate. If the noise level is smaller than a threshold value, such a generalization is regarded as a reasonable one. Otherwise, the generalization is contradictory.

Furthermore, a rule in the GDT-RS is selected according to its priority. The priority can be defined by the number of instances covered (matched) by a rule (i.e. the more instances are covered, the higher the priority is), by the number of attributes occurring on the left-hand side of the rule (i.e. the fewer attributes, the higher the priority is), or by the rule strength [12].

4.1.2 Searching Algorithm for an Optimal Set of Rules:

Searching algorithm developed by Dong *et al.* [13] for a set of rules and based on the GDT-RS methodology is outlined below.

Step 1. Create the GDT.
Step 2. Calculate the probabilities of generalizations.
Step 3. For any compound instance *u'* (such as the instance u'_1 in the above table), let $d(u')$ be the set of the decision classes to which the instances in *u'* belong.
Step 4. Using the idea of the discernibility matrix, create a discernibility vector (i.e. the row or the column with respect to *u* in the discernibility matrix) for *u*.
Step 5. Compute the entire local relative reducts for instance *u* by using the discernibility function.
Step 6. Construct rules from the local reducts for instance *u*, and revise the strength of each rule using (4).
Step 7. Select the best rules from the rules (for *u*) obtained in *Step 6* according to its priority [12].
Step 8. $U' = U' - \{u\}$. If $U' \neq \emptyset$;, then go back to *Step 4*. Otherwise, go to *Step 9*.
Step 9. If any rule selected in *Step 7* covers exactly one instance, then Stop, otherwise, repeat the above steps to select a minimal set of rules covering all instances in the decision table.

4.2 Rough sets with heuristics (RSH)

Rough set with heuristics (RSH) is a system that helps to select attribute subset [16]. The development of the RSH is based on the following observations: (i) a database always contains a lot of attributes that are redundant and not necessary for rule discovery; (ii) if these redundant attributes are not removed, not only does the time complexity of the rule discovery increase, but also the quality of the discovered rules can be significantly decreased. The goal of attribute selection is to find an optimal subset of attributes according to some criterion so that a classifier with the highest possible accuracy can be induced by an inductive learning algorithm using information about data available only from the subset of attributes.

4.2.1 Heuristic Algorithm for Feature Selection:

The attributes from database set called *CORE are* as an initial attribute subset. Next, attributes were selected one by one among the unselected ones using some strategies, and add them to the attribute subset until a reduct approximation is obtained.

Algorithm

Let *R* be a set of selected condition attributes, *P* a set of unselected condition attributes, *U* a set of all instances, and *EXPECT* an accuracy threshold. In the initial state, let $R = CORE(C)$, $P = C - CORE(C)$ and $k = 0$.

Step 1. Remove all consistent instances:
$U = U - POS_R(D)$.
Step 2. **If** k _ *EXPECT* **then** *STOP*
else if $POS_R(D) = POS_C(D)$,
return *`only k is available'* and
STOP
Step 3. Calculate v_p , m_p.
Step 4. Choose the best attribute p, i.e. that with the largest $v_p \times m_p$, and set $R = R \cup \{p\}$, $P = P - \{p\}$
Step 5. Go back to *Step 2*.

4.3 Rough sets and boolean reasoning (RSBR)

RSBR is a system for discretization of real-valued attributes. Discretization of real valued attributes is an important preprocessing step in the rule discovery process. The development of RSBR is based on the following observations: (i) real-life data sets often contain mixed types of data such as real-valued, symbolic data, etc. (ii) real-valued attributes should be discretized in preprocessing (iii) the choice of the discretization method depends on the analyzed data.

The main module in the rule discovery process is the GDT-RS. In the GDT-RS, the probabilistic distribution between possible instances and possible generalizations depends on the number of the values of attributes. The rules induced without discretization are of low quality because they will usually not recognize new objects.

4.3.1 Discretization based on RSBR:

In order to solve the discretization problems, a discretization system called the RSBR was developed which is based on hybridization of rough sets and Boolean reasoning. [17], [18]

A great effort has been made [19], [20] to find effective methods of discretization of real-valued attributes. Different results may be obtained by using different discretization methods. The results of discretization affect directly the quality of the discovered rules. Some of discretization methods totally ignore the effect of the discretized attribute values on the performance of the induction algorithm. The RSBR combines discretization of real-valued attributes and classification. In the process of the discretization of real-valued attributes it should also take into account the effect of the discretization on the performance of the induction system GDT-RS.

Roughly speaking, the basic concepts of the discretization based on the RSBR can be summarized as follows: (i) discretization of a decision table, where $Vc = (v_c, w_c)$ is an interval of real values taken by attribute c, is a searching process for a partition P_c of V_c for any $c \in C$ satisfying some optimization criteria (like a minimal partition) while preserving some discernibility constraints [17, 18] (ii) any partition of V_c is defined by a sequence of the so-called *cuts* $v_1 < v_2 < \text{---} < v_k$ from V_c (iii) any family of partitions $\{P_c\}$ $c \in C$ can be identified with a set of cuts.

4.4 Map reduce method

Map Reduce Method is a programming model [21]. It is capable of processing large data sets called Big Data. It is well suited to handle in a distributed computing environment of real world tasks. The Map reduce computation is described in the fig.3 by two function as follows,

Map Function

It takes input pair (attribute/value) and produces set of intermediate attribute/value pairs.

Reduce Function

It accepts intermediate attribute/value pairs, merges to form smaller sets of values. Typically 0 or 1 output value produced per reduce invocation.

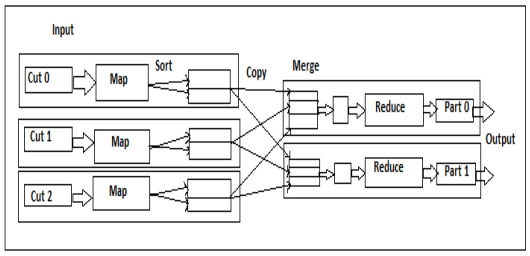


Fig.3. Map Reduce Programming Model

4.4.1 Applying Map Reduce Model to Compute Rough Set Approximation*:*

From the given Information System, the Universal set is first partitioned into a number of subsets [22]. From the subsets Equivalence Classes are obtained in a single step using Map Function. Now these equivalence classes are combined if it derives the same information set with respect to their conditional attributes. Similarly Equivalence classes of different sub decision classes / tables can be combined together if their information set is same. The above steps are executed in parallel.

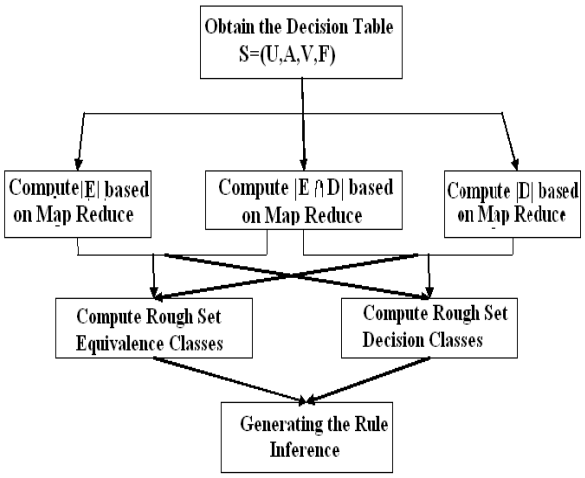


Fig.4. An example of calculating Rough Set Approximations based on Map Reduce

As Equivalence (E) and Decision classes (D) are computed, construction of association between these two classes can be done in parallel as shown in fig. 4. If there is a relation between the two classes, then association exist, otherwise no. Then lower and upper approximation indexes are computed directly, leading to the calculation of lower and upper approximation.

These parallel methods run on different clusters namely Hadoop, Phoenix, and Twister [23]. Among these Twister is faster than other two systems and Hadoop is slower than other two. Users can decide which runtime system to be used in their application.

4.4.2 Dynamic Data Mining:

Variation in the Data plays a vital role in recent years. Hence it is necessary to dynamically update knowledge as given in fig.5. The knowledge updating takes place under the variation of object set, attribute set and attribute value.

Knowledge Updating due to Variation of Object Set

Here two situations arise, there may be single or multi object enters in/gets out of the information system [24]. Former is called Immigration and latter called Emigration of object. These processes reflect the refining of knowledge in neighborhood decision table.

The following steps help to understand the above said cases. Due to arrival of new single/multi objects, there may or may not be new decision classes generated.

Step 1: The decision classes are updated.
Step 2: Neighborhoods of Immigration object is computed.
Step 3: Neighborhoods of the universe are updated.
Step 4: Finally lower, upper approximation of the decision classes are updated.
Similar to Immigration, in the Emigration of Single/Multi objects also, there may or may not be deletion of existing decision classes. The steps involved are same as above.

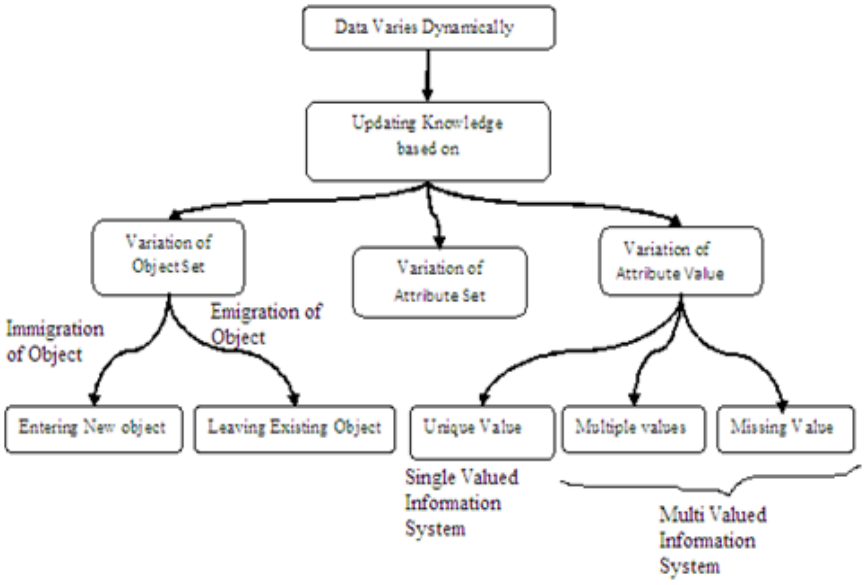


Fig.5. Dynamic Data Interpretation

Knowledge Updating due to Variation in Attribute Value-Matrix Based Approaches

Attribute value in the information system may be represented either by single value or multi value even sometimes may be missed [25]. If an attribute has unique value, then the information system

is called single valued Information system. If an attribute has multi value, then the Information system is called multi valued Information system. If there is missing value, it can also be treated as multi valued Information system referred in fig.5.

Static / Dynamic Algorithm for updating approximations under adding / deleting objects steps as follows.

When the object enters in / gets out of the decision table,

Step 1: Relation matrix is updated.
Step 2: Induced Diagonal matrix is updated.
Step 3: Decision matrix is updated.
Step 4: Intermediate matrices or cut matrices are also updated.
Step 5: Lower and Upper approximations are generated.
Step 6: Finally probabilistic positive, boundary and negative approximations are updated.

Incremental Algorithm-Matrix Based Approaches

Static / Dynamic Algorithm [26] for updating approximations under adding / deleting objects is given below.
When the object enters in / gets out of the decision table,
Step 1: Relation matrix is updated.
Step 2: Induced Diagonal matrix is updated.
Step 3: Decision matrix is updated.
Step 4: Intermediate matrices or cut matrices are also updated.
Step 5: Finally lower and upper approximations are generated.

5. CONCLUSION

This paper attempts to bring out the concepts of rough set theory applied to data mining. Basic concepts of the data mining and the various steps normally employed were discussed. In traditional methods of decision making, scientific expertise in combination with some statistical methods is used to support the management. But these cannot be used to handle big data. This paper discusses the various techniques employed by the researchers in identifying the vagueness in data bases using rough set theory. The potential applications of rough set theory in data mining are reviewed in this paper. The concepts of many recent data mining techniques using rough set theory such as GDT-RS, RSH, RSBR, Map Reduce, Dynamic Data mining are also consolidated and presented in the above sections.

REFERENCES

[1] Pawlak, Z., "Rough Sets", International Journal of Information and Computer Sciences, Vol. 11, pp. 341- 356, 1982.
[2] Pawlak, Z., "Rough Sets: Theoretical Aspects of Reasoning about Data", Kluwer Academic Publishers, ISBN 0-79231472, Norwell-USA, 1991.
[3] Pawlak Z, Grzymala-Busse J, Slowinski R, and Ziarko W, "Rough Sets", Communications of the ACM, Vol. 38, No. 11,pp 89-95,Nov 1995.
[4] Zarandi M.H.F, Kazemi A, "Application of Rough Set Theory in Data Mining for Decision Support Systems (DSSs)", Journal of Industrial Engineering, Vol. 1, pp 25 – 34, 2008.
[5] Mert Bal, "Rough Sets Theory as Symbolic Data Mining Method", An Application on Complete Decision Table Information Science Letters, Vol. 2, pp. 35-47, NSP Natural Sciences Publishing, 2013.

[6] Pawlak Z, "Rough Classification", International Journal of Man-Machine Studies, Vol. 20, No. 5, pp. 469-483, 1984.

[7] Grzymala-Busse, J.W., "LERS-A System for Learning from Examples Based on Rough Sets", Slowinski, R., (Ed.) Intelligent Decision Support Handbook of Application and Advances of the Rough Sets Theory, Kluwer Avademic Publishers, 1992.

[8] Stefanowski, J., "The Rough Set Based Rule Induction Technique for Classification Problems", Proceedings of 6th European Conference on Intelligent Techniques and Soft Computing, EUFIT 98, Aachen, Germany, pp.109-113, 1998.

[9] Mienko R., Stefanowski, J., Taumi, K.& Vanderpooten, D., "Discovery-Oriented Induction of Decision Rules", Cahier du Lamsade, No.141, Université Paris Dauphine,1996.

[10] Grzymala-Busse, J.W., Stefanowski, J., "Three Discretization Methods for Rule Induction", International Journal of Intelligent Systems, Vol. 16, pp. 29-38, 2001.

[11] Stefanowski, J., Vanderpooten, D., "Induction of Decision Rules in Classification and Discovery-Oriented Perspectives", International Journal of Intelligent Systems, Vol. 16, pp. 13-27, 2001.

[12] Zhong N., Dong J.Z. and Ohsuga S, "Data mining: A probabilistic rough set approach", Rough Sets in Knowledge Discovery, Vol.2, Heidelberg, Physica-Verlag, pp.127-146, 1998.

[13] Dong J.Z., Zhong N. and Ohsuga S, "Probabilistic rough induction: The GDT-RS methodology and algorithms", Foundations of Intelligent Systems, Berlin, Springer, pp.621-629, 1999.

[14] Komorowski J., Pawlak Z., Polkowski L. and Skowron A, "Rough sets: A tutorial, In Rough Fuzzy Hybridization", A New Trend in Decision Making, Singapore, Springer, pp.3-98, 1999.

[15] Pawlak Z. and Skowron A, "A rough set approasch for decision rules generation", Proceedings Workshop W12: The Management of Uncertainty in AI at 13th IJCAI, pp.1-19, 1993.

[16] Dong J.Z., Zhong N. and Ohsuga S, "Using rough sets with heuristics to feature selection", New Directions in Rough Sets, Data Mining, Granular-Soft Computing, Berlin, Springer, pp.178-187, 1999.

[17] Nguyen H. Son and Skowron A, "Quantization of real value attributes", Proceedings International Workshop Rough Sets and Soft Computing, 2nd Joint Conference on Information Sciences (JCIS'95), Durham, NC, pp.34-37, 1995.

[18] Nguyen H. Son and Skowron A, "Boolean reasoning for feature extraction problems", Foundations of Intelligent Systems, Berlin, Springer, pp.117-126, 1997.

[19] Fayyad U.M. and Irani K.B. (1992) "On the handling of real-valued attributes in decision tree generation", Machine Learning, Vol.8, pp.87-102. 1992.

[20] Nguyen H. Son and Nguyen S. Hoa, "Discretization methods in data mining", Rough Sets in Knowledge Discovery, Heidelberg, Physica-Verlag, pp.451-482, 1998.

[21] Zhang J, Li T, Pan Y, "Parallel Rough Set Based Knowledge Acquisition Using Map Reduce from Big Data", ACM, Beijing, China, August 12, 2012.

[22] Zhang J, Li T, Ruan D, Gao Z, Zhao C, "A parallel method for computing rough set approximations", Information Sciences, Elsevier Inc, Vol.194, pp 209–223, 2012.

[23] Zhang J, Wong J.S, Lia T, Pan Y, "A comparison of parallel large-scale knowledge acquisition using rough set theory on different map reduce runtime systems", International Journal of Approximate Reasoning, Elsevier, 2013.

[24] Zhang J, Li T, Ruan D, Liu D, "Neighborhood Rough Sets for Dynamic Data Mining", International Journal of Intelligent Systems, Wiley Periodicals, Inc.,Vol. 27, pp 317–342, 2012.

[25] Zhanga J, Li T, Ruan D, Liud D, "Rough sets based matrix approaches with dynamic attribute variation in set-valued information systems", International Journal of Approximate Reasoning, Elsevier Inc.,Vol. 53, pp 620–635, 2012.

[26] Zhang J, Li T, Chen H, "Composite rough sets for dynamic data mining", Information Sciences, Elsevier Inc., Vol. 257, pp 81–100, 2013

AUTHORS

Selvi received her B.E., degree from Madras University, Chennai, India in 1998. She also received her M.E., degree from Anna University, Chennai, India in 2007. She is currently a PhD candidate at the Faculty of Information and Communication Engineering, Anna University Chennai, India. She has got 14 Years of Teaching Experience. Now she is working as Assistant Professor at the department of Computer Science and Engineering, Government College of Engineering, Bargur , Tamilnadu, India from 2013. She is interested in Data Mining, Cloud Computing, Soft Computing, Network Security and Wireless Sensor Network.

R.Priya,
B.E Final Year,
Department of Computer Science and Engineering,Government college of Engineering-Bargur,
Tamilnadu,
India.

V.Anitha,
B.E Final Year,
Department of Computer Science and Engineering,
Government college of Engineering-Bargur,
Tamilnadu,
India.

V.Divya Bharathi,
B.E Final Year,
Department of Computer Science and Engineering,
Government college of Engi neering-Bargur,
Tamilnadu,
India.
.

AUTHOR INDEX