Proceedings ICETCSE 2016

3rd International Conference on Emerging Technologies in Computer Science & Engineering

17th and 18th October 2016
V. R. Siddhartha Engineering College,
Vijayawada, India

Published & Edited by

International Journal of Computer Science and Information Security (IJCSIS)

Vol. 14 Special Issue ICETCSE 2016

ISSN 1947-5500

Pennsylvania, USA

Indexed and technically co-sponsored by :

Editorial
Message from Editorial Board

*It is our great pleasure to present the **ICETCSE 2016 Special Issue** (Volume 14) of the **International Journal of Computer Science and Information Security** (IJCSIS). High quality research, survey & review articles are proposed from experts in the field, promoting insight and understanding of the state of the art, and trends in emerging technologies and computer science. It especially provides a platform for high-caliber academics, practitioners and PhD/Doctoral graduates to publish completed work and latest research outcomes. According to Google Scholar, up to now papers published in IJCSIS have been cited over 6827 times and the number is quickly increasing. This statistics shows that IJCSIS has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. There have been many improvements to the processing & indexing of papers; we have also witnessed a significant growth in interest through a higher number of submissions as well as through the breadth and quality of those submissions. IJCSIS is indexed in major academic/scientific databases and important repositories, such as: Google Scholar, Thomson Reuters, ArXiv, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, ResearchGate, Academia.edu and EBSCO among others.*

*On behalf of IJCSIS community and the sponsors, we congratulate the authors, the reviewers and thank the committees of **3rd International Conference on Emerging Technologies in Computer Science & Engineering (ICETCSE 2016) from V. R. Siddhartha Engineering College, Vijayawada, India,** for their outstanding efforts to review and recommend high quality papers for publication. In particular, we would like to thank the international academia and researchers for continued support by citing papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status.*

*"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." For further questions or other suggestions please do not hesitate to contact us at **ijcsiseditor@gmail.com**.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 14, Special Issue ICETCSE 2016 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(ijcsiseditor@gmail.com)
Pennsylvania, USA
Tel: +1 412 390 5159

Proceedings of the 3rd International Conference on Emerging Technologies in Computer Science & Engineering

ICETCSE-2016

October 17-18, 2016

(Sponsored by TEQIP - II S.C. 1.2)

Organized by:

DEPARTMENTS OF

Computer Science and Engineering
Information Technology

VELAGAPUDI RAMAKRISHNA SIDDHARTHA ENGINEERING COLLEGE

IJCSIS/ICETCSE EDITORIAL BOARD

Editorial Board Members
Prof. Dr. Jacek M. Czerniak, Casimir the Great University in Bydgoszcz, Poland
Dr K. Suvarna Vani, Professor, Dept of CSE, VR Siddhartha Engineering College
Prof. Dr. Binh P. Nguyen, National University of Singapore
Dr S. Vasavi, Professor, Dept of CSE, VR Siddhartha Engineering College
Professor Seifeidne Kadry, American University of the Middle East, Kuwait
Prof. Dr. Riccardo Colella, University of Salento, Italy
Prof. Dr. Sedat Akleylek, Ondokuz Mayis University, Turkey
Prof. Dr. Sachin Kumar, Indian Institute of Technology (IIT) Roorkee
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Dr. Naseer Alquraishi, University of Wasit, Iraq
Dr. Kai Cong, Intel Corporation, & Computer Science Department, Portland State University, USA
Dr. Omar A. Alzubi, Al-Balqa Applied University (BAU), Jordan
Dr. Shimon K. Modi Director of Research BSPA Labs, Purdue University, USA
Professor Ying Yang, Computer Science Department, Yale University, USA

ICETCSE 2016 List of Papers

(4) Research Scholar, MGR University, Maduravoyal, Chennai, India

ICETCSE 2016 Paper 18: Speed Optimised AES-GCM (pp. 34-37)
N Rajitha (#), R Sridevi (*)
(#) Research Scholar, Department of Computer Science & Engineering, JNTUHCEH, Hyderabad, Telangana, India
(*) Professor, Department of Computer Science & Engineering, JNTUHCEH, Hyderabad, Telangana, India

ICETCSE 2016 Paper 19: Telugu numeral recognition using machine learning techniques (pp. 38-45)
P. Harish, II M.Tech, Computer Science & Engineering, VR Siddhartha Engineering College, Kanuru, Vijayawada, Andhra Pradesh, India.
Dr. S. Vasavi, Professor, Computer Science & Engineering, VR Siddhartha Engineering College Kanuru, Vijayawada, Andhra Pradesh, India

ICETCSE 2016 Paper 22: Exploring Spectral Features for Emotion Recognition Using GMM (pp. 46-52)
J. Naga Padmaja (1), R. Rajeswar Rao (2)
(1) Computer Science and Engineering, JNTU Hyderabad, Khammam, INDIA
(2) Computer Science and Engineering, JNTU Vizianagaram, Vizizanagaram, INDIA

ICETCSE 2016 Paper 29: Perceptive Reckon System with RFID-tag with IEEE 802.15.4 technology (pp. 53-56)
Azeem Mohammed Abdul (#1), Syed Umar (#2)
(#1) M.tech Student, Department of Electronics and Communication Engineering, KL University, Vaddeswaram.
(#2) Professor, Department Of Computer Science Engineering, GIST, Jaggayyapet

ICETCSE 2016 Paper 30: Efficient Mining of Top-K High Utility Itemsets from Uncertain Databases (pp. 57-62)
Vamsinath Javangula (1), Suvarna Vani Koneru (2), Haritha Dasari (3)
(1) CSE Department, P.B.R V I TS, Kavali, Andhra Pradesh, India
(2) CSE Department, Velagapudi Ramakrishna Siddhartha Engineering College, Kanuru, Andhra Pradesh, India
(3) CSE Department, University College of Engineering College, Kakinada, JNTUK, Andhra Pradesh, India

ICETCSE 2016 Paper 34: Enhanced Caesar Cipher Algorithm with Variable Length Key and Increased Cipher Complexity (pp. 63-68)
P. Srinivasa Rao (1), Dr. D. Nagaraju *(2)

(1) Research Scholar, Dept. of CSE, Acharya Nagarjuna University, Nagarajuna Nagar, Guntur, A.P, India.
(2) Dept. of Information Technology, Lakireddy Balireddy College of Engineering, Mylavaram, A.P, India.

ICETCSE 2016 Paper 36: Student Performance Analysis Using Educational Data Mining (pp. 69-76)
P Ramya, Gudlavalleru Engineering College, Gudlavalleru, Krishna (Dt), Vijayawada
M Mahesh Kumar, LakiReddy BaliReddy College of Engineering, Mylavaram, Krishna (Dt), Vijayawada

ICETCSE 2016 Paper 37: Detecting and Preventing CSRF Attack on Web Application (pp. 77-80)
N. Vidya Rani , Dr. G. Ramakoteswara rao
Department of Information technology,
VR Siddhartha Engineering College,
Vijayawada, Andhra Pradesh, India

ICETCSE 2016 Paper 38: Review on Bastion Hosts (pp. 81-88)
G. Vijayababu (1), D. Haritha (2), R. Satya Prasad (3)
(1) Gate Faculty, Jeevan Engineers Academy, Guntur,Andhra Pradesh, INDIA
(2) Professor, S.R.K. Institute of Technology, Vijayawada, Andhra Pradesh, INDIA
(3) Associate Professor, Acharya Nagarjuna University, Guntur, Andhra Pradesh, INDIA

ICETCSE 2016 Paper 39: A unique dimensionality shrinkstylefor highdimensional spatiotemporal brain signal data based on graph signal processing theory (pp. 89-95)
Dr. D. Nagaraju *(1), A. Sarvani (2), B. Venugopal (3)
(1,2) Dept. of Information Technology, Lakireddy Balireddy College of Engineering, Mylavaram, A.P, India.
(3) Dept. of CSE, Andhra Loyola Institute of Engineering and Technology, Vijayawada, A.P, India.

ICETCSE 2016 Paper 40: Support Vector Machine Based Classification for Face Recognition (pp. 96-100)
D Sudha Rani #1 ,P. Swathi #2, V Srinivasa Rao #3 ,K Srinivas #4
Department of Computer Science and Engineering, VRSiddhartha Engineering College, Vijayawada, Andhra Pradesh, India.

ICETCSE 2016 Paper 41: Classification with Active Learning Method in Relevance Feedback for Content-Based Image Retrieval (pp. 101-105)

Suresh Tommandru, Dr. D. Naga Raju
Dept. of Information Technology, Lakireddy Bali Reddy College of Engineering, Mylavaram, A.P, India.

ICETCSE 2016 Paper 42: Protein Secondary Structure Extraction using Bag of Words Model (pp. 106-110)

K. Sushma, Department of Computer Science and Engineering, V. R. Siddhartha Engineering College (VRSEC), Andhra Pradesh, India.
K. Suvarna Vani, Department of Computer Science and Engineering, V. R. Siddhartha Engineering College (VRSEC), Andhra Pradesh, India

ICETCSE 2016 Paper 43: Analysis of Breast Cancer Diagnosis using Cytological Images (pp. 111-115)

O. likhitha, Department of Computer Science and Engineering, V. R. Siddhartha Engineering College (VRSEC), Andhra Pradesh, India
K. Suvarna Vani, Department of Computer Science and Engineering, V.R. Siddhartha Engineering College (VRSEC), Andhra Pradesh, India.

ICETCSE 2016 Paper 48: Analysis of Different Bioinformatics Analytic Procedures in Biomedical Big Data Evaluation (pp. 116-123)

P. Udayaraju (1) , Dr. K. Suvarnavani (2) , Dr. Chandra Sekhar Vasamsetty (3)
(1) Assistant Professor Department of CSE, SRKR Engineering College, Bhimavaram, AP, India
(2) Professor, Department of CSE, VR Siddhartha Engineering College, Kanuru, Vijayawada, AP, India
(3) Associate Professor Department of CSE, SRKR Engineering College, Bhimavaram, AP, India

ICETCSE 2016 Paper 53: Multilingual Text Categorization (pp. 124-130)

Nadella. Haritha, Dr. M. Suneetha
Department of Information and Technology, VR Siddhartha Engineering College (VRSEC), Vijayawada, Andhra Pradesh, India

ICETCSE 2016 Paper 57: A Study on Meta Data Extraction Systems and Features of Cloud Monitoring (pp. 131-135)

S. Amarnadh (1)*, V. Srinivasa Rao (1), M.A. Rama Prasad (1), V. Venkateswara Rao (2)
(1) Department of Computer Science and Engineering, Chirala Engineering College, Chirala-523 157
(2) Mathematics Division, Department of Science and Humanities, Chirala Engineering College, Chirala-523 157

ICETCSE 2016 Paper 58: Managing Disaster Event using Geospatial and Web Technologies (pp. 136-139)

G. Rajasekhara Basava Kumar (#), Nitin Mishra (*), G. Srinivasa Rao (*), V. Bhanumurthy (*), J. V. D. Prasad (#)
(#) Department of Computer Science, VRSEC Vijayawada 520 007, Andhra Pradesh, India
(*) Remote Sensing Applications Area, National Remote Sensing Centre-ISRO Balanagar, Hyderabad 500 037, Telangana, India

ICETCSE 2016 Paper 59: Configuration Monitoring and Auditing of LAN Switches (pp. 140-145)

G. Krishna Kishore, D. S Lakshmi, M. Narasimha
CSE-DCCT ITD, VRSEC-DRDL, kanuru Vijayawada India-Kanchanbagh Hyderabad India

ICETCSE 2016 Paper 61: Impact of Location Popularity on Throughput and Delay in Mobile Ad Hoc Networks (pp. 146-151)

Dr. G. Krishna Kishore (1), R. Navya (2), Rajesh K (3)
(1) Dept of CSE, VR Siddhartha Engineering College
(2) Dept of CSE, VR Siddhartha Engineering College
(3) Dept of CSE, DVR & Dr. HS MIC College of Technology

ICETCSE 2016 Paper 64: Comparative Analysis of Shadow Detection and Removal Methods on an Image (pp. 152-156)

V. Rashmi (1), V. Srinivasa Rao (2), K. Srinivas (3)
(1) PG Scholar, Dept. Of CSE, V. R. Siddhartha Engineering College
(2) Professor & Head, Dept. Of CSE, V. R. Siddhartha Engineering College
(3) Professor, Dept. Of CSE, V. R. Siddhartha Engineering College

ICETCSE 2016 Paper 65: Comparative Analysis of Machine Learning Techniques on Stock Market Prediction (pp. 157-161)

M. Sreemalli (1), P. Chaitanya (2), K. Srinivas (3)
(1) Dept. of CSE, V.R Siddhartha Engineering College
(2) Dept. of CSE, V.R Siddhartha Engineering College
(3) Professor, Dept. of CSE, V.R Siddhartha Engineering College

ICETCSE 2016 Paper 66: A Relative Study on Open Source IaaS Cloud Computing Tools (pp. 162-166)

Bala Savitha Jyosyula, Suhasini Sodagudi
Department of Information Technology, VRSEC, Vijayawada, Andhra Pradesh, India.

ICETCSE 2016 Paper 67: Ear Biometrics System based on Gray Level (Spatial) Statistical Feature Extraction (pp. 167-171)

P. Ramesh Kumar, Department of Computer Science & Engineering, V.R. Siddhartha Engineering College, Vijayawada- 520 007, INDIA
SS Dhenakaran, Department of Computer Science & Engineering, Alagappa University, Karaikudi - 630 003, INDIA

ICETCSE 2016 Paper 68: Person Re-Identification across Multiple Camera View (pp. 172-176)
V Ramya , V V Vineela, V Srinivasa Rao, K Srinivas,
Department of Computer Science and Engineering, VR Siddardha Engineering College, Vijayawada, Andhra Pradesh, India.
* Software Group, Advanced Data Processing Research Institute– ISRO, Hyderabad, Telangana, India.

ICETCSE 2016 Paper 69: A Comprehensive Survey on Big Data Analytics and Techniques (pp. 177-184)
K. Naresh Babu, Asst.prof, Dept of IT, Geethanjali college of engineering and technology, Hyderabad , India,
Dr. Suneetha Manne, Prof and HOD, Dept of IT, Velagapudi, RamakrishnaSiddhartha Engineering college, Vijayawada, India,

ICETCSE 2016 Paper 70: GIS the Future of Utility Management (pp. 185-188)
P Sree Gayathri (1), V Phani Kumar (2)
(1) PG Scholar, Dept. of CSE, V.R Siddhartha Engineering College
(1) Asst Professor, Dept. of CSE, V.R Siddhartha Engineering College

ICETCSE 2016 Paper 71: An Algorithm for Basic IoT Architecture (pp. 189-193)
Rajesh Vemulakonda (#1), Sowjanya Meka (*2), Venkatesh Ketha (#3), Phani Praveen Surapaneni (#4), Sri Vijaya Kondapalli (*5)
(#) Computer Science & Engineering Department, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, India
(*) Department of information Technology, Prasad V. Potluri Siddhartha Institute of Technology Vijayawada, India

ICETCSE 2016 Paper 74: Dynamic Search for Spatio-Textual Queries on Location Based Applications (pp. 194-198)
K. Haritha (1), M. Vani Pujitha (2*)
(1) PG Scholar, Dept. of CSE, V.R Siddhartha Engineering College
(2*) Assistant Professor, Dept. of CSE, V.R Siddhartha Engineering College

ICETCSE 2016 Paper 76: Heart Disease Diagnosis Using Predictive Data Mining (pp. 199-204)
M. Swathi Lakshmi, Dr D. Haritha
SRKIT, Vijayawada, India

ICETCSE 2016 Paper 77: Healthcare Applications of the Internet of Things (IoT): A Review (pp. 205-213)
M.V.D.N.S. Madhavi (1), K. Hemalatha (1); P.V.S. Sairam (2), D. Rajani (1)
(1) Department of Mathematics, V.R.Siddhartha Engineering College, Andhra Pradesh, India
(2) Department of Physics, Andhra Loyola College, Vijayawada

ICETCSE 2016 Paper 82: Prescriptive Analytics for Intelligent Business Systems (pp. 214-220)
Raghuvira Pratap A, J V D Prasad, Kranthi Kumar G, Dr. Suvarna Vani K
V.R. Siddhartha Engineering College

ICETCSE 2016 Paper 83: Implementing Power Distribution System Using Geographic Information System (pp. 221-224)
A. Sowmya (1), A. Jitendra (2)
(1) PG Scholar, Dept. of CSE, V.R Siddhartha Engineering College (VRSEC).
(2) Professor, Dept. of CSE, V.R Siddhartha Engineering College (VRSEC)

ICETCSE 2016 Paper 84: A Study on Social Engineering Attacks and Defence Mechanisms (pp. 225-231)
Mukesh Chinta (#1), Jitendra Alaparthi (#2), Eswar Kodali (#3)
(#1, #2, #3) Department of CSE, V R Siddhartha Engineering College Vijayawada, Andhra Pradesh, India

ICETCSE 2016 Paper 86: A Survey on Big Data Analysis for High Velocity Data (pp. 232-237)
K.S.Vijayalakshmi, Asst. Professor: CSE Department, VRSEC, Vijayawada, India
Vamsi Nadella, Grad Student: University of Georgia, Athens, Atlanta, USA
Dr. K. V. Sambasiva Rao, Dean: CSE Department, NRI Institute of Technology.
Dr. E. V. Prasad, Director: LBR College of Engineering.
Dr. V. Srikanth, Director: Citi Bank, UK

ICETCSE 2016 Paper 62: Comparative Study Review on Lung Cancer Detection Using Optimization and Clustering Approach (pp. 238-243)
Divya Chauhan, M. Tech., Shoolini University, Solan, India (H.P.)
Varun Jaiswal, Assistant Professor, Shoolini University, Solan, India (H.P.)

Image Quality Assessment for Land use and Land Cover Classes using SRKR Model

Nalluri Sunny[1]
Computer Science and Engineering
V. R. Siddhartha Engineering College
nallurisunny4848@gmail.com

Dr. V. Srinivasa Rao[2]
Computer Science and Engineering
V. R. Siddhartha Engineering College
drvsrao9@gmail.com

Srikanth Mithinti[3]
Computer Science and Engineering
V. R. Siddhartha Engineering College
srikanth.mithinti@gmail.com

Srikanth Chitturi[4]
Computer Science and Engineering
V. R. Siddhartha Engineering College
chsrikanth90@gmail.com

***Abstract*—Images can suffer with distortions due to several sources, from the acquisition process itself to compression, noisy channels and so on. Images can also undergo quality improvement processes, in water marking, compression, restoration, synthesis, satellite images, signal acquisition, storage, re-construction, authentication, presentation and reproduction or restoration techniques. In every case it is useful to quantify the quality of such resulting image. In order to identify the land use and land cover classes a low resolution (LR) MODIS satellite image is taken as input, but the highest resolution (HR) of MODIS image is only 250 meters per pixel which is not suitable for the identification of land cover classes. Hence, for that LR satellite image if quality is calculated then it will be easier to identify them. Once the quality is calculated then it is magnified, de-blurred and it is subjected through kernel regression algorithm. A bi -cubic interpolation is applied to the LR image in order to get magnified and de-blurred HR image. The resulting de-blurred HR image is applied to K-means clustering algorithm to get the particular land cover classes.**

***Index Terms*— Gradient similarity, Moderate Resolution Imaging Spectroradiometer (MODIS), Super Resolution Kernel Regression (SRKR), High Resolution and Low Resolution Image, K-Means Clustering.**

I. INTRODUCTION

Image quality assessment is made with Gradient Similarity by comparing two image blocks in any two images. Here low resolution image and blurred image from Moderate Resolution Imaging Spectroradiometer (MODIS) [1,2,3] are taken as input and image quality is calculated and then it is subjected to bi-cubic interpolation in order to get de-blurred image. MODIS data play a predominant role in detecting human activities across the world. As land covers i.e., the surface of the earth changes seasonally MODIS data will be more effective in identifying them. Generally MODIS data is used to detect wild fires, deforestation, floods [3] etc.. The land cover types are divided into 4 types. They are (1) the land with human activities, (2) the land without human activities, (3) the water without ice, (4) the land cover without snow and ice.

A Kernel regression based super-resolution algorithm (SRKR) [4] is used in order get the de-blurred HR image. The goal of super-resolution (SR) is to estimate a high resolution (HR) image from one or a set of low-resolution (LR) images The SRKR only uses a single low resolution image as the input and generates a higher resolution image by a de-blurring process with up-sampling. Then we apply Super Resolution Kernel Regression (SRKR) algorithm to synthesize the high resolution MODIS image from its original low resolution version, 250m per pixel. The constructed super-resolution image can achieve 4 to 8 times higher resolutions, which can avoid the resolution limitation on land activity detected in the satellite image. In order to identify the land cover types K-means clustering algorithm is used.

II. PROPOSED APPROACH

The proposed approach consists of two modules. Image quality assessment for low resolution MODIS image, bi-cubic interpolation is applied to that LR image and applying K-means clustering in order to identify land cover types from that quantified image. Image quality is made by the gradient similarity method (with multiple Masking Parameter's K' values) which is shown below

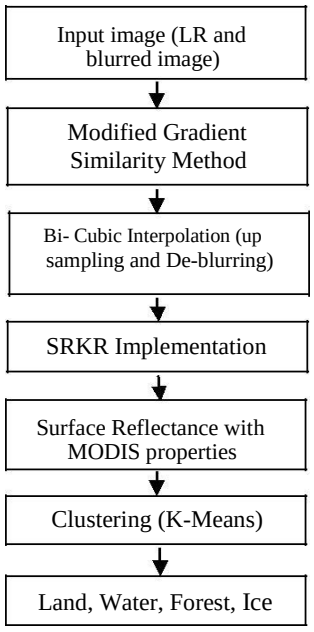


Fig. 1. Block Diagram of the proposed

A. Modified Gradient Similarity Method

In this method block wise or pixel wise comparison of images is made in order to calculate image quality [5]. Block wise comparison is most suitable method to calculate image quality because in pixel wise comparison the gradient values for the same group of pixels changes according to the position of these pixels. The overall gradient similarity should lie between 0 and 1 only. The contrast and structural changes in an image gives the gradient similarity.

The gradient similarity is defined as [5]

$$g(x,y)=\frac{2(1-R)+K}{1+(1-R)^2+K}$$

Where R is the masked gradient change which is given by [5]

$$R=\frac{\left|g_x-g_y\right|}{\max\left(g_x-g_y\right)}$$

Masked Gradient, change gives the amount of distortion that is overlapped on the blurred image from the de-blurred image[5]. Where

$$K=\frac{C_4}{\max\left(g_x-g_y\right)}$$

In order to avoid the numerator being zero C4 is taken as a small constant (10^{-5})

There are 2 problems with the above "Equation (1)". They are

(I). In a distorted image all parts of the image may or may not have distortions so when the block where the distortion is not present is considered then the gradient difference between the original image block and the distorted image block will be similar i.e., R=0. In this situation the gradient similarity will be completely dependent on constant K value.

(II). If R = 1 then | g_x - g_y | = max(g_x , g_y) i.e., overall image quality is increasing than 1 i.e.; false gradient is created.

In order to avoid these problems the K value should be modified as [5]

$$K=\frac{K`}{\max(g_x-g_y)}$$

Where K' is a positive constant and it is called as a masking parameter. The range between the terms 2(1- R) and 1 + (1- R)^2 lies between [20,40] . The choice of K' lies between 200 and 1000 because on an average the lies between max (g_x, g_y) lies between 50 and 5.

On Substituting the K value in g(x,y), we get [5]

$$g(x,y)=\frac{2(1-R)+K^{'}/\max\left(g_x,g_y\right)}{1+(1-R)^2+K^{'}/\max\left(g_x,g_y\right)}$$

The proposed approach describes about

As K' value lies between 200 and 1000, at a regular interval the Masking parameter (K') values are chosen as 200,300,400,500,600,700,800,900 and 1000 and then the gradient similarity is calculated. Though these K' values are increased the image quality lies between 0 and 1 only.

B. Measurement of Luminance Distortion

In addition to contrast and structural changes, luminance changes also must be considered, though they have less impact on the image quality assessment [5]. Luminance is an indicator of how bright the surface of an image will appear. A squared pixel error method is used in order to calculate luminance similarity. Luminance similarity is given by [5]

$$e\left(x_i,y_i\right)=1-\left(\frac{x_i-y_i}{L}\right)^2$$

Where xi, and Yi are the pixels at position I in image blocks x and y respectively, and L is the dynamic range of pixel values i.e., 255 for 8 bit grayscale images.

C. Adaptive Distortion integration

Gradient and luminance similarities are integrated in order to derive the overall image quality indicator q(x_i ,y_i) is given by[5]

$$q=\left(1-W\left(g,e\right)\right).g+W\left(g,e\right).e$$

Where q , g , e are the abbreviated forms of q(x_i ,y_i) , g(x_i ,y_i), e(x_i, y_i) respectively, and W(g, e) is the weighting function which is calculated by [5]

$$W(g,e)=p\,.\,g$$

Where p is a positive weighting parameter which is taken as 0.1 because luminance similarity has less impact on image quality when compared to structural and contrast changes.

The resultant LR image quality is judged and it is subjected to method of bi-cubic interpolation in order to magnify the image and to remove blurriness.

D. Bi-Cubic Interpolation

Bi-cubic interpolation [6] considers 16 pixels (4x4). Images which are re-sampled with Bi-cubic interpolation is smoother and have fewer interpolation artifacts. Interpolation is a method of constructing new data points within the range of a discrete set of known data points [7]. A different problem which is closely related to interpolation is the approximation of a complicated function of a simple function and Up-sampling is the process of increasing the sampling rate of a signal.

E. SRKR Implementation

The SRKR algorithm consists of two major steps: (a) up-sampling and (b) de-blurred as shown in the below figure 2(a). Up-sampling is the process of increasing the sampling rate of a signal. In the up-sampling step [8], we perform bi-cubic interpolation of the input low resolution image (LR) with a desired scaling factor. Then the low resolution image and the interpolated high resolution image are partitioned into corresponding pixels. After Kernel Regression (KR)

[9,10,11,12] algorithm is applied on the obtained targets and the feature vectors. The KR model is then used to predict the blurred high resolution (HR) image using the bi-cubically interpolated high resolution image as its input feature vector.

In the de-blurring step, we further blur and down-sample the blurred HR image to obtain the blurred LR image. Similar to the up-sampling step, the blurred HR image, the blurred LR image, and the original LR image is partitioned into pixels as shown in Figure 2(b). For each pixel of the original LR image, the pixels are sampled as training targets, and the neighboring pixels in the corresponding blurred LR patch are taken as feature vectors for each sampled pixel. The KR model is then used to predict the de-blurred HR image using the neighboring pixels of the blurred HR image.

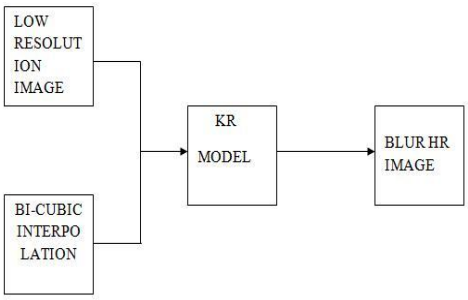


Fig. 2(a) Up Sampling

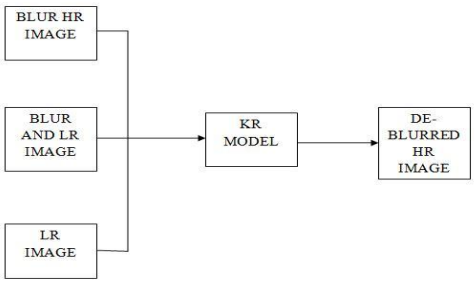


Fig. 2(b) De-blurring

F. Surface Reflectance

For each selected pixel, surface reflectances of a patch are extracted in the super-resolution MODIS image. The patch width is the scaling factor multiplied by 4. By multiplying by the scaling factor 4 the resultant image is more magnified and the blurriness is removed. The maximum pixel range to classify the land cover types is taken from MODIS data [4].

TABLE I. MINIMUM AND MAXIMUM AREA RANGE VALUES FROM MODIS DATASET

Minimum	Maximum	Area
0	150	Ocean
151	190	Land
191	220	Land with Forest
221	225	Land with Snow

Table I shows the information about the minimum and the maximum surface reflectance MODIS data set. These covers the landscapes like an ocean, bare land, land with forest and land with snow. Based on these surface reflectance values the K-Means clustering algorithm classifies into the respective landscapes.

G. K- Means Clustering

K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters [13]. This procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) for a fixed a priority [14]. The maximum pixel range to classify "Forest", "Land", "sea", and "Ice" land cover types according to their surface reflectance are taken from the MODIS dataset which are mentioned in table 1.

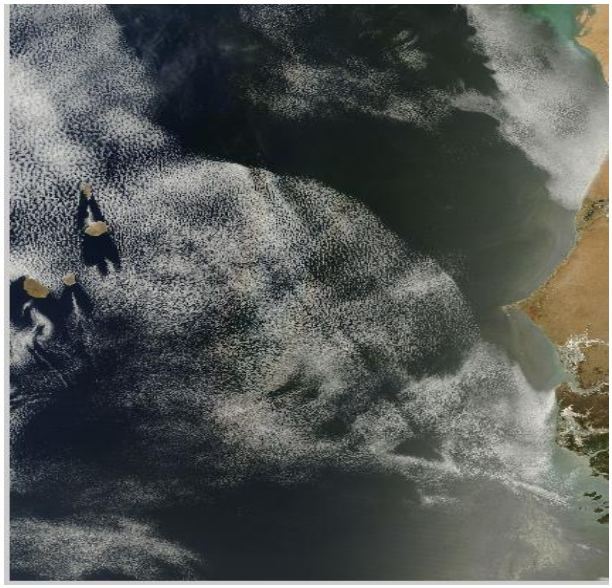


Fig. 3(a) Original LR image

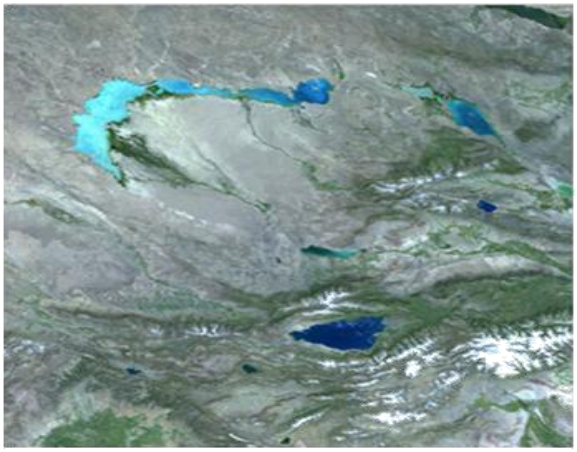


Fig. 3(b) Resultant de-blurred image

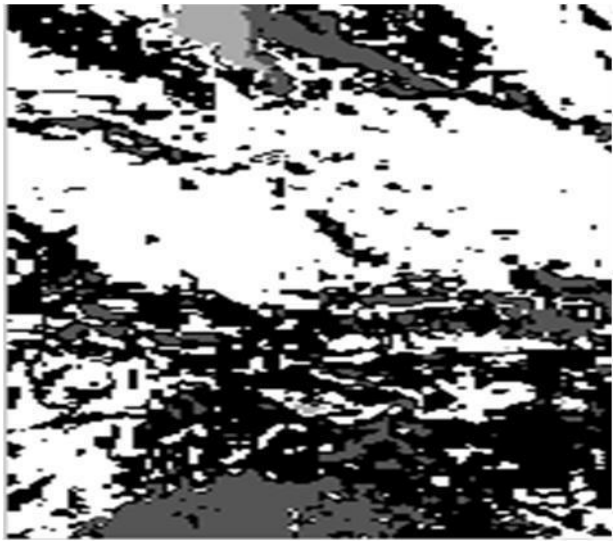
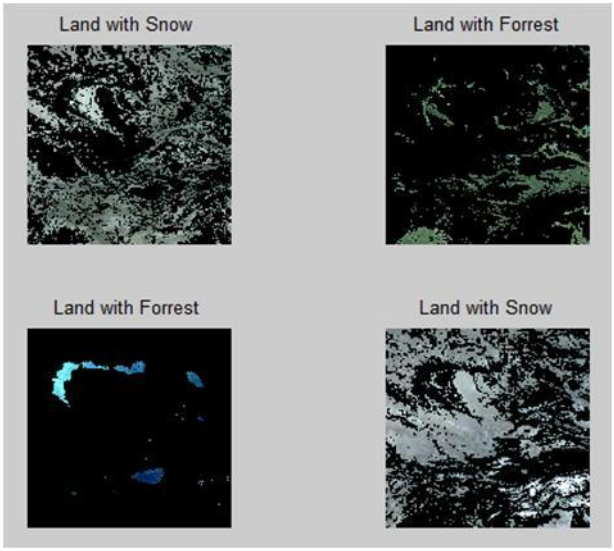


Fig. 4: Super-resolution through the SRKR algorithm: (a) Clustered Image (Black, White, Gray)



(b) Required land cover classes

Figure. 3(a) shows the original LR image, to which the quality is assessed, then interpolation techniques are applied to get de-blurred image, Figure. 4(a) shows the clustered image through SRKR algorithm and Figure 4(b) shows the resultant land cover classes based on the surface reflectance values of MODIS dataset.

III. Results

In this paper image quality is calculated by the modified gradient similarity method with multiple masking parameter (K') values and this quantified image is subjected to interpolation techniques and further SRKR algorithm is applied in order to classify the land cover classes forest, land, sea, ice by K- means clustering [15].

Table II: Image Quality for different databases when Masking parameter(K') increases

Masking Parameter (K' Values)	Final Image Quality				
	LIVE Database	CSIQ Database	TID Database	A57 Database	IVC Database
200	0.9980	0.9246	0.9982	0.9991	0.9963
300	0.9986	0.9490	0.9984	0.9992	0.9976
400	0.9989	0.9614	0.9986	0.9993	0.9982
500	0.9991	0.9690	.99888	0.9994	0.9985
600	0.9992	0.9741	0.9990	0.9995	0.9988
700	0.9993	0.9777	0.9992	0.9996	0.9989
800	0.9994	0.9805	0.9994	0.9997	0.9991
900	0.9995	0.9826	0.9996	0.9998	0.9992
1000	0.9996	0.9843	0.9997	0.9999	0.9993

Table III: Evaluation Criteria when higher value of K' value is considered.

Evaluation Criteria	LIVE		CSIQ		TID	
	Before	After	Before	After	Before	After
SROCC	0.9554	0.9618	0.9126	0.9319	0.8554	0.8857
KROCC	0.8131	0.8299	0.7403	0.7553	0.6651	0.7098

Table II represents the final image quality for different K' values for different types of database images, i.e., for highest K' value the image quality is increased. As we can see that the image quality is higher when K' value is 1000. Table III represents the evaluation criteria before and after i.e., before means when K' value is taken as 200 and after means when maximum K; value is considered.

IV. CONCLUSION

By modifying the masking parameter (K') values the quality of the image is improved. For experimental evaluation, we used various types of images from 6 image databases which are Laboratory for Image and Video Engineering (LIVE) database [16], Categorical Image Quality (CSIQ) database [17], Tampere Image Database (TID) [18], Toyama [19], A57 [20] and IVC [21] databases and 3 evaluation criteria which are Spearman's Rank Correlation Coefficient (SROCC) [22], Kendall's Rank Correlation Coefficient (KRCC) [22], Pearson Linear Correlation Coefficient (PLCC) [22]. When compared to previous techniques, only human activities are observed, but by using SRKR algorithm, various types of land cover classes are easily identified.

V. REFERENCES

[1] R. Fries, M. Hansen, J. Townshend, R. Sohlberg, "Global land cover classifications at 8 km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers", International Journal Remote Sensing, 1998.

[2] M. Pilloni, M. Melis, and A. Marini, "Analysis and validation of a methodology to evaluate land cover change in the mediterranean basin using multitemporal MODIS data", Present Environment and Sustainable Development, 2010.

[3] X. Zhan, R. Sohlberg, J. Townshend, C. DiMiceli, M. Carroll, J. Eastman, M. Hansen, and R. DeFries, "Detection of land cover changes using MODIS 250m data", Remote Sensing of Environment 83, 336-350, 2000.

[4] H. He and W. Siu, "Single image super-resolution using Gaussian process regression", Computer Vision and Pattern Recognition, 2011

[5] Image Quality Assessment Based on Gradient Similarity Anmin Liu, Weisi Lin, *Senior Member, IEEE,* and Manish Narwaria vol 21, April , 2012.

[6] F. Fekri, R. M. Mersereau, and R.W. Schafer, "A generalized interpolative VQ method for jointly optimal quantization and interpolation of images," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 5, 1998.

[7] J. W. Woods, "Two-dimensional Kalman filters ," in Two-Dimensional Digital Signal Processing , T. S. Huang, Ed. New York: Springer-Verlag, 1981.

[8] H. A. Aly and E. Dubois. Image up-sampling using total variation regularization with a new observation model. IEEE Trans. on IP, 14(10):1647–1659, 2005

[9] K. Teknomo, Kernel regression, available at regression, 2007.

[10] K. Teknomo, Kernel regression, available at: http://peopl e.revoledu .com/kardi/tutorial /regression/kernel regression, 2007.

[11] G.Watson, Smooth regression analysis, *The Indian Journal of Statistics*, Series A 26 (1964), 359-372.

[12] A. Nadaraya, On estimating regression, Theory of Probability and its Applications 9 (1964), 141-142.

[13] S. Ray, R.H. Turi, "Determination of number of clusters in K-means clustering and application in colthe image segmentation", Proc. 4th ICAPRDT, pp. 137-143, 1999.

[14] [14] J.L. Marroquin, F. Girosi, "Some Extensions of the K-Means Algorithm for Image Segmentation and Pattern Classification", Technical Report, MIT Artificial Intelligence Laboratory, 1993.

[15] K. Atsushi, N. Masayuki, "K-Means Algorithm Using Texture Directionality for Natural Image Segmentation", IEICE technical report. Image engineering, 97 (467), pp.17-22, 1998

[16] H. R. Sheikh, K. Seshadrinathan, A. K. Moorthy, Z.Wang, A. C. Bovik, and L. K. Cormack, "Image and video quality assessment research at LIVE," [Online]. Available: http://live.ece.utexas.edu/research/quality/

[17] E. C. Larson and D. M. Chandler, "Categorical image quality (CSIQ) database," [Online] Available: http://vision.okstate.edu/csiq.

[18] N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, and V. Lukin, "Metrics performance comparison for color image database," in Proc. 4th Int. Workshop Video Process. Quality Metrics Consum. Electron., Scottsdale, AZ, Jan. 2009.

[19] Y. Horita, K. Shibata, Y. Kawayoke, and Z. M. P. Sazzad, "MICT image quality evaluation database, "[Online] Available:http://mict.eng.u-toyama.ac.jp/mict/index2.html.

[20] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," [Online]. Available: http:// foulard.ece.cornell.edu /dmc27/vsnr/vsnr.html

[21] A. Ninassi, P. Le Callet, and F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding," in *Proc. SPIE: Human Vis. Electron. Imag.*, San Jose, CA, Jan. 2006, vol. 6057.

[22] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[23] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[24] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[25] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[26] K. Elissa, "Title of paper if known," unpublished.

[27] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[28] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[29] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Cloud Sharing for Economic Benefits

Yallamanda Challa[1], V Gopinath[2], K Purna Prakash[3] and Dr S Krishna Rao[4]
Department of Information technology
Sir C R Reddy college of Engineering, Eluru
[1]challa.ynaidu@gmail.com, [2]velivela.gopi@gmail.com, [3]prakash.nani@gmail.com and [4]skrao71@gmail.com

Abstract- Now a day's cloud computing is more efficient in current world technologies. Cloud computing provide Software as a Service, Platform as a Service and infrastructure as a Service. In this client has to pay money for the applications or software`s for limited usage and limited period (it may be number of days or number of months). Here we are proposing that, share our service or license to the trusted third-party; they can use application and services as in client-cloud available. Like our mobile having the application like Shareit we can share application without download from play store and we can use it, same way the trusted third party user can get the service from the existing client as their benefit. Even we can give as license out for co-existing client to use our client-cloud benefits. To share the cloud we may get proper authentication for the existing client cloud owner. It's like a small scale business to overcome the economical benefits from the cloud server. It enhances the various applications and software's with sharing benefits.

Keywords: *Client-cloud, Sharing cloud, Authentication, Trusted third party*

I. INTRODUCTION

Cloud computing provides computation, software, data access, and storage services that do not require end-user knowledge of the physical location and configuration of the system that delivers the services. Parallels to this concept can be drawn with the electricity grid, wherein end-users consume power without needing to understand the component devices or infrastructure required to provide the service. Cloud computing describes a new supplement, consumption, and delivery model for IT services based on Internet protocols, and it typically involves provisioning of dynamically scalable and often virtualized resources. It is a byproduct and consequence of the ease-of-access to remote computing sites provided by the Internet.

This may take the form of web-based tools or applications that users can access and use through a web browser as if the programs were installed locally on their own computers. Cloud computing providers deliver applications via the internet, which are accessed from a web browser, while the business software and data are stored on servers at a remote location. In some cases, legacy applications (line of business applications that until now have been prevalent in thin client Windows computing) are delivered via a screen-sharing technology, while the computing resources are consolidated at a remote data center location; in other cases, entire business applications have been coded using web-based technologies such as AJAX. Most cloud computing infrastructures consist of services delivered through shared data-centers and appearing as a single point of access for consumers' computing needs. Commercial offerings may be required to meet service-level agreements (SLAs), but specific terms are less often negotiated by smaller companies. Cloud computing technologies can offer important benefits for IT organizations and data centers running MTC applications: elasticity and rapid provisioning, enabling the organization to increase or decrease its infrastructure capacity within minutes, according to the computing necessities.

II. EXISTING SYSTEM

Cloud computing is a kind of Internet-based computing that provides shared processing resources and data to computers and other devices on demand. It is a model for enabling ubiquitous, on-demand access to a shared pool of configurable computing resources, which can be rapidly provisioned and released with minimal management effort. Cloud computing and storage solutions provide users and enterprises with various capabilities to store and process their data in third-party data centers. It relies on sharing of resources to achieve coherence and economy of scale, similar to a utility over a network.

III. PROPOSED SYSTEM

Proposed system we are organizing the cloud computing with economical benefits oriented. In this, we proposed the sharing cloud concept to reduce the economical burden from the client. Here shared cloud can get the proper authentication from the client cloud. Shared cloud has the proper credentials from the client to accept the client cloud with time basis or a license based. Based on this

we proposed that sharing cloud is very much economic benefits for the existing client cloud.

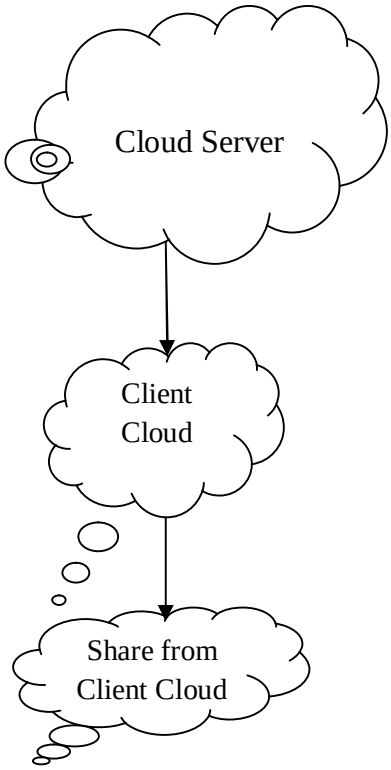


Figure 1: Sharing cloud architecture

IV. LITERATURE SERVEY

A. Cloud computing characteristics:

1. Agility improves with users' ability to re-provision technological infrastructure resources.

2. Application programming interface (API) accessibility to software that enables machines to interact with cloud software in the same way the user interface facilitates interaction between humans and computers. Cloud computing systems typically use REST-based APIs.

3. Cost is claimed to be reduced and in a public cloud delivery model capital expenditure is converted to operational expenditure. This is purported to lower barriers to entry, as infrastructure is typically provided by a third-party and does not need to be purchased for one-time or infrequent intensive computing tasks. Pricing on a utility computing basis is fine-grained with usage-based options and fewer IT skills are required for implementation (in-house).

4. Device and location independence enable users to access systems using a web browser regardless of their location or what device they are using and accessed via the Internet, users can connect from anywhere.

5. Multi-tenancy enables sharing of resources and costs across a large pool of users thus allowing for: Centralization of infrastructure in locations with lower costs.

6. Peak-load capacity increases (users need not engineer for highest possible load-levels)

7. Utilization and efficiency improvements for systems that are often only 10–20% utilized.

8. Reliability is improved if multiple redundant sites are used, which makes well-designed cloud computing suitable for continuity and disaster recovery.

9. Performance is monitored and consistent and loosely coupled architectures are constructed using web services as the system interface.

10. Security is often as good as or better than under traditional systems, in part because providers are able to devote resources to solving security issues that many customers cannot afford.

B. Cloud services

Cloud services promise great benefits in terms of financial savings, easy and convenient access to data and services, as well as business agility. Organizations and individuals therefore choose to outsource their data to the cloud, where an un-trusted party is in charge of storage and computation. A major concern for the adoption of cloud computing is the inability of the cloud to build user trust in the information security measures deployed in cloud services. Common computing techniques cannot be applied on encrypted data, and therefore the data and the programs that compute on the data must be decrypted before being run on the cloud infrastructure. A comprehensive solution for securing the cloud computing infrastructure can be based on cryptographic mechanisms of secure computation. These mechanisms allow for distributed computation of arbitrary functions of private (secret) inputs, while hiding any information about the inputs to the functions. Put differently, these mechanisms support computation on encrypted data.

C. Motivation

Cloud Computing has moved beyond the peak of inflated expectations and will be widely adopted by companies in about two to five years. Due to the anticipated advantages of Cloud Computing, as e. g. high flexibility and

costs many companies do not analyze their decisions carefully. This approach raises risk factors like for instance hidden Costs or a vendor-lock-in which discrete the pursued benefits. Thus, companies should conduct an ex ante analysis of direct and indirect costs to mitigate certain risk factors and to be aware of important cost types and factors. In this paper we present a cloud sharing from the existing cloud server to get the economical befit. The main focus of our model lies in the identification and calculation of cost factors. The results serve as decision support by evaluating Cloud Computing Services and providers on a cost basis. We based our model on the analysis of real Cloud Computing Services from our Cloud Computing research data base. The sharing cloud model is prototypically implemented on a website for further evaluation steps and is accessible for the general public. The software tool is able to analyze the cost structure of Cloud Computing Services and thus supports decision makers in validating Cloud Computing Services from a cost perspective. The presented multi-method approach extends the sharing cloud and applies deductive and inductive methods to develop a theoretically and practically based model.

V. SYSREM DESCRIPTION

(i) Sharing Cloud

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service-provider interaction". To incorporate the majority of possible costs and cost categories of Cloud Computing Services we applied a cloud sharing techniques approach. While traditional accounting approaches primarily aim at identifying the lowest possible costs, the benefits of the sharing approach lie in the improvement of customer-supplier communication and the analysis of the whole lifecycle of cloud computing usage. Furthermore, the sharing cloud approach makes it possible to analyze the costs or individual cost components of a client cloud by means of a predefined scheme.

- Transparency: We provide an in-depth description of the model and the applied criteria.
- Applicability: The prototypical implemented software tool allows for an easy application of the sharing model with reasonable effort.
- Variability: The sharing model is standardized largely, but central aspects are variable, so that desired changes or extensions of the model are possible.

(ii) *Economic Impact of Cloud Computing*

Cloud computing likely will extend the IT induced economic growth in developed economics and foster growth in economies where IT penetration is not yet fully mature. We conclude that governments should work together to take advantage of the benefits of cloud computing.

The new big thing of the IT world is "cloud computing", a general purpose technology that could provide a fundamental contribution to promote efficiency in the private and public sectors and promote growth, competition, and business creation.

Cloud computing is an Internet-based technology (hence "cloud") which stores information in servers and provides it as an on-demand service. The economic impact of cloud computing will be substantial on both households and companies.

On one side, consumers will be able to access all of their documents and data from any device (the home or work PC, the mobile phone, an internet point), as they already do for email services or social networks. On the other side, firms will be able to rent computing power (both hardware and software in their latest versions) and storage from a service provider, while paying on demand, as they already do for other inputs such as energy and electricity.

(iii) Economic Impact of Cloud Computing on Business

Cloud computing will allow firms to rent software and data storage remotely on an as-needed basis. This should reduce the fixed costs firms incur when they enter a new market or begin production of a good. Other economics benefits of cloud computing includes:

(i) Rapid software updates and easy modification of software
(ii) Cost sharing among consumers
(iii) Variable computational capacity and increased efficiency

(iv)Energy savings as servers are moved to cold climates

As firms become cheaper to start and operate, the European Union can expect to see several hundred thousand new firms develop across its member states. Cloud computing could also result in lower prices for consumers, increased competition among firms, and an increase in European Union output growth by several tenths of a percentage point. This is significant; Gains in the short term will be much greater if member governments support adoption through financial incentives, general promotion,

and agreements to reduce barriers to data transmission across borders.

(v) Oblivious user management for cloud

One of the main issues with data sharing in cloud environment is to manage user access and its auto revocation in a controlled and flexible way. The issue becomes more complex when privacy on user access has to be ensured as well to hide additional leakage of information. For automatic revocation over cloud data, access can bounded within certain anticipated time limit so that the access expires beyond effective time. This time-oriented approach is more rigid and not a one-size-fits-all solution. In certain circumstances, exact time anticipation is not an easy choice. Instead, the alternate solution could be task oriented to restrict user beyond certain number of permissible attempts to access the data. We have proposed oblivious user management (OUM) in which a user can have access on cloud data for certain number of attempts without imposing any time restriction. For user authorization and her subsequent revocation, owner will perform one time setup activity and that is same for all users. The model also alleviates the burden of managing different access parameters at user end with each request, as she will always use the same parameter for all valid attempts. Our approach also conceals the privacy of user attempts throughout the communication. Hiding this information helps to avoid distinguishing importance of particular user that has more authorization over others.

(vi)Computation with multiple servers

Cloud services provide a powerful resource to which weak clients may outsource their computation. While tremendously useful, they come with their own security challenges. One of the fundamental issues in cloud computation is: how does a client efficiently verify the correctness of computation performed on an un-trusted server? Furthermore, how can the client be assured that the server learns nothing about its private inputs? With the ever-increasing usage of cloud computation in practice, it is not surprising that the area of verifiable computation, that addresses precisely these questions, has generated widespread interest in the cryptography community. In recent years, a number of proposals have been made for constructing verifiable computation protocols.

(vii) Cloud sharing privacy

That guarantee privacy of inputs (in addition to the correctness of computation) relies on the use of fully homomorphism encryption (FHE). An unfortunate consequence of this dependence on FHE is that all hope of making verifiable computation implementable in practice hinges on the challenge of making FHE deployable in practice.

VI. RELATED WORK

Cloud computing providers may simplify pricing models. However, until that day, when determining total pricing for a cloud computing service, consider the services offered by the provider and how each pricing structure is worded. Using more than one cloud has its benefits, but can also get expensive. To manage costs, examine your providers' data access charges, as well as your application design. The cloud, it's a term thrown around in casual conversation these days. Microsoft's ad campaign directed “To the cloud!” Apple crams iCloud down your throat at every opportunity; billboards on my commute say nothing more than “Cloud Computing" on them. Most people don't even know what it is or what it means, and frankly they don't care, it just sounds cool. To the layman, the cloud is synonymous with the internet itself, and if they're talking about SaaS products, they're basically right. To developers dealing with cloud computing it's more complicated however. To be clear, I'm talking about public cloud computing in terms of infrastructure.

a. Amazon – AWS
b. Microsoft – Azure
c. Google - Cloud Platform
d. RackSpace – OpenStack

When you are starting out a new SaaS project you generally have some minimal requirements. You will need a web server and probably a database server. When you are just getting your application off the ground, it will have a low amount of users that you hope to grow over time. The number of users will determine the load on your server and the server specs will determine how many users are too many. So where do you start? For startups, I believe there are two broad approaches to building an application and the infrastructure to run it on:

A) - Pessimistic - Build it quickly, get it out there, and validate the business before spending the time to engineer it for scaling.

B) - Optimistic - Build it carefully, code for scalability, and launch it with the assumption that it must scale quickly.

With method A, your primary goal is to get the thing off the ground and hope that scaling becomes a concern. If you start to exhaust the resources on your initial

server, you're probably gaining traction in the market. You can buy yourself some time by scaling vertically (adding more RAM, CPU power etc.) while you work on re-engineering your code for horizontal scaling. With method B, you're sold on the viability of this product and you know it's going to be a hit. You believe that any hiccups in service would be more damaging than coming to market later. Therefore, you're going to spend the time to code the software for scalability up front and configure your infrastructure for growth.

VII. Conclusion

In this paper we are concluding that cloud sharing for the trusted third party, they can use application and services as in client-cloud available it same way the trusted third party user can get the service from the existing client as their benefit. Here existing client cloud get benefit (In case of money) forms the sharing cloud member. By providing the services in a license based, client-cloud trusts the third-party user. If the use using more the than his/her premises client-cloud can stop authentications and ask them to redo the process of getting authentication. So the economically cloud user get the benefit from the shared users. However the process of getting prior permission from the cloud server then only we access and we can stop using services.

References

[1] P.S. Barreto, H. Y. Kim, B. Lynn, and M. Scott, "Efficient algorithms for pairing-based cryptosystems," in *Proc. Crypto*, vol. 2442, *Lecture Notes in Computer Science*, 2002, pages 354–368.

[2] R. Bhatti, J. Joshi, E. Bertino, and A. Ghafoor, "X-GTRBAC Admin: A decentralized administration model for enterprise-wide access control," in *Proc. ACM SACMAT*, 2004, pages 78–86.

[3] M. Blaze, J. Feigenbaum, and A. D. Keromytis, "KeyNote: Trust management for public-key infrastructures," in *Proc. Security Protocols Int. Workshop*, 1998, pages 59–63.

[4] M. Blaze, J. Feigenbaum, and J. Lacy, "Decentralized trust management," in *Proc. IEEE Symp. Security Privacy*, May 1996, pages 164–173.

[5] D. Boneh and M. K. Franklin, "Identity-based encryption from the Weil pairing," in *Proc. CRYPTO*, vol. 2139, *LNCS*, 2001, pages 213–229.

[6] D. Boneh, C. Gentry, B. Lynn, and H. Shacham, "Aggregate and verifiably encrypted signatures from bilinear maps," in *Proc. Eurocrypt*, 2003, pages 416–432.

[7] D. Boneh, C. Gentry, B. Lynn, and H. Shacham, "A survey of two signature aggregation techniques," *CryptoBytes*, vol. 6, no. 2, pages 1–9, 2003.

[8] D. Boneh, B. Lynn, and H. Shacham, "Short signatures from the Weil pairing," in *Proc. Asiacrypt*, vol. 2248, *LNCS*, 2001, pages 5 14–532.

[9] D. Clarke, J.-E. Elien, C. Ellison, M. Fredette, A. Morcos, and R. L. Rivest, "Certificate chain discovery in SPKI/SDSI," *J. Comput. Security*, vol. 9, no. 4, pages 285–322, Jan. 2001.

[10] P. Devanbu, M. Gertz, C. Martel, and S. Stubblebine, "Authentic third-party data publication," *J. Comput. Security*, vol. 11, no. 3, pages 291–314, 2003.

[11] John Rhoton, "Cloud Computing Explained: Implementation Handbook for Enterprises", 2015.

[12] David S. Linthicum, "Cloud Computing and SOA Convergence in Your Enterprise: A Step-by-Step Guide", vol no. 3, 2004.

[13] George Reese, "Cloud Application Architectures: Building Applications and Infrastructure in the Cloud", 2013.

[14] Tim Mather," Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance", 2011.

[15] Andy Mulholland, "Enterprise Cloud Computing: A strategy Guide for Business and Technology Leaders", 2014.

[16] Yallamanda Challa, "An Identity key based exchange protocol", page no. 3, 2012.

[17] Federico Etro, "Source Review of Business and Economics", Vol. 54, No. 2, page no. 179-208, 2009.

[18] George Reese, "Building Applications and Infrastructure in the Cloud", Cloud Application Architectures, 2010.

NoSQL for Census Data Analysis

M.Akka Lakshmi
Lakshmi.muddana@gitam.in

G.Victor Daniel
victordaniel.gera@gitam.in

D.Srinivasa Rao
srinu.dhanionda@gitam.in

ABSTRACT*: Data is growing exponentially with Terra bytes of data being generated daily by social networks, millions of mails. Voluminous data is being generated by enterprises in the form of documents. Meaningful insights can be derived from this huge amount of data that helps organizations improve their business. 'NoSQL' technologies provide solution with high performance and scalable approach to analyze large and non-structured datasets. In this paper, we analyze Census data ,using various 'NoSQL' technologies to gain insights into workforce available in various states and age groups of India as per Census-2011.

Keywords- Big data, Non-structured data, NoSQL

I INTRODUCTION

The amount of data produced by mankind from beginning of time till 2003 is only 5 billion gigabytes but the same amount was created in every 2 days in 2011 and every 10 min in 2013. It is estimated that by 2020 the total data will be about 35ZB with China accounting for more than $1/5^{th}$ of it. It is also estimated that about $1/3^{rd}$ of all data will exists or pass through cloud by 2020.

Knowingly or unknowingly every one of us are contributing to this voluminous data. There are more than billion internet users generating zetabytes of internet traffic every day in the form of millions of e-mails, millions of blogs and hundreds of websites being created every minute. Users upload 48 hours of new video every minute. Social networks like Face book, Twitter, LinkedIn are also contributing terra bytes of data daily. In addition, business organizations generate and accumulate lot of documents and other transactional data.

Traditionally, organizations store and preserve only structured data and is used for analysis to take business decisions. But this constitute small portion of total data , about 10 to 15%. Enterprise data represent large percentage of text and multimedia data which is in the form of semi-structured or un-structured. And also of the total data, about 85% to 90% is non-structured and cannot be processed by traditional database systems. Analyzing small portion of data may not provide precise insights to observe trends in the data sets. Large portions of data in unstructured or semi-structured form is unused. Usually organizations retain this data for specific period of time and then disposed off . If this data is also considered for analysis, it leads to more accurate results. Analysis of large data sets gives more meaningful long term co-relations among data sets and help organizations in their business. Such accurate analyses help organizations in improving their business which will in turn generate more revenue.

In this paper , section II discuss the need of non-relational databases and popular NoSQL databases, in section III we analyze the census data using Hive and Neo4J and section IV is conclusion.

II Non-Relational databases

A. Need of Non-Relational databases increased the use of Internet. Developments in web technologies, proliferation of social networks are contributing to exponential growth of data. Most of this data is in non-structured form. Relational databases cannot handle data beyond certain size and can only process data that is structured.

The following challenges of relational model have driven the emergence of new data models. (i) impedance mismatch- the way data is represented in relational databases is different from the way it is represented in memory (ii) scalability- as the data size grows, we have to scale-out rather than scale-up (iii) single point failures. This demands for a technology that can store and process large datasets in structured, non-structured form and at the same time giving good performance, scalability and avoiding single point failures. This has given rise to development of NoSQL databases.

B. Number of technological advancements and user needs have contributed to the development of NoSQL technologies for analyzing big data sets (i) Dramatic decrease in storage costs- storage capacity is roughly doubled every 40 months since 1980s[2]. Organizations started storing and retaining the data for long periods (ii) Increase in processing power. (iii) emergence of data centers and cloud computing that provide flexibility for storage and computing (iv) current workloads demand scale-out and not scale-up(v) todays data is large and unstructured. Many proprietary and open source NoSQL databases have emerged. The popular being ,Googles Big table, Facebook's Cassandra, Amazon's DynamoDB, MongoDB, HBase.

C.Big data analytics

NoSQL databases are suited for Big Data analytics. Big data is characterized by (i) large Volume (ii) Variety in terms of data sources and data types like structured, semi-structured and unstructured data, data generated by machines, networks (iii) Velocity- the pace at which data is being generated. Atomic reactors generate 40TB per sec, 640TB of data is generated for one flight. (iv) Veracity & Validity -Data to be analyzed should be clean and valid for the given application to make appropriate decisions. Organizations should keep away dirty data from being accumulated.

Big data analytics find its applications in many areas like predicting customer behavior, health care, developing security systems for crime and fraud detection, remote sensing data for weather prediction, agriculture yield estimation to achieve food security.

Popular use cases of big data technologies include (i) areas of security for intrusion and fraud detection , developing spam filters (ii) resource optimizations for internet and other organizational resources(iii) medical data analysis to predict and prevent spread of deceases, predicting patient readmissions (iv) manufacturing sector to determine optimal time for repair or replacement of machines (v) retail sector for building recommendation systems , posting relevant Ads to customers (v) Agriculture to estimate and predict yield, Land Cover and usage patterns to improve farming practices to achieve food security.[3]

In the big data analysis, social networks like Face book, Twitter, LinkedIn plays a key role for utilizing information and messages posted by users. It is used by (i) Customer product companies to know product feedback, Product defects, user preferences, Customer behavior. (ii) Advertising and Marketing agencies to understand responses to their campaigns and promotions (iii) Sports teams to track ticket sales, know team strategies (iv) to predict Election results Processing [7] and also expressing graph computations.[5]

III Data analysis

Census- 2011 data pertaining to workers in India has been analyzed, using Hadoop Hive and Neo4J, with the following results

TABLE I. workers in India

Workers Category	Male	Female	Rural	Urban
Main workers	75.35%	24.65%	67.81%	32.19%
Marginal workers	49.22%	50.78%	86.22%	13.78%
Non-workers	39.97%	60.04%	66.53%	33.47%

Pre-processing: The following Duplicate data is deleted .Age like 60+ is deleted since it is already included in 60-69 and 70-79 age groups.30-59 age groups have been deleted as it is included in in more specific age groups. From age group 40 onwards range of 10 was given but for age groups 5 to 29 , range of 5 was given. Since it is workers data, analysis was done from age 20 onwards. Hence age group 20-24 and 25-29 is merged to form a group of 10 range and also 30-34 and 35-39 are also merged. Analysis is as follows. Overall population distribution is as given below

TABLE II. Analysis of data given in various Age-groups starting from 5 years to 80+

Male population	Female Population	Rural Population	Urban Population
51.47%	48.53%	68.86%	31.14%

NoSQL databases – key-value stores, Document databases, and Column family databases are based on aggregate model. Data set often contains relationships or connectedness among the data. Relational databases can model data relationships but not efficient when the relationships are many. Graph data structure is the natural way of representing relationships using edges or links. We find number of applications of graphs in the areas of social networks, computer networks, transport networks, protein-protein interactions. Neo4J is the popular graph databases. Census data is analyzed and represented as graph depicts the states having maximum and minimum percent of Main, Marginal and Non-workers which is drawn using Neo4J. The nodes are state names and category of significant domination of main workers compared to Non-workers till age of 60. Good number of main workers can also be found in the age group 60-69

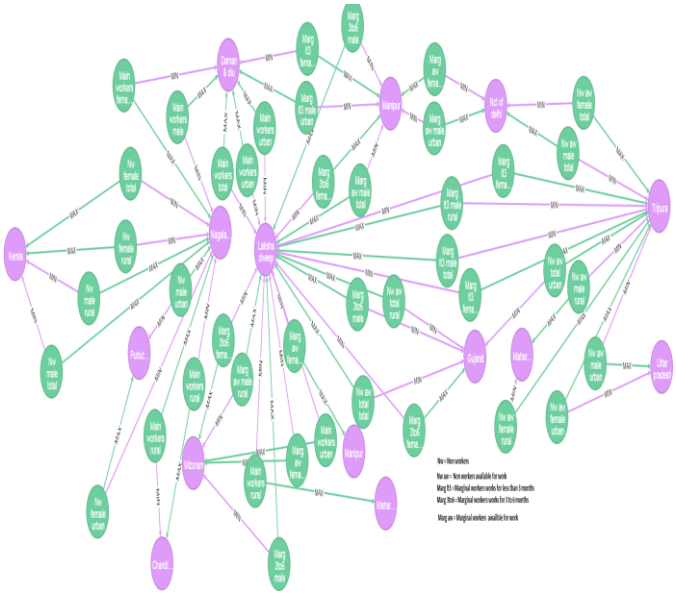


Figure. I state wise maximum and minimum percent of Main, Marginal and Non-workers

We can see more Non-workers than Main workers in the age group of 20-29. Later, we can find significant domination of main workers compared to Non-workers till age of 60. Good number of main workers can also be found in the age group 60-69.

Workers data is with further divided into male, female, rural and urban of main, marginal and non-workers .Rural population is more than double the urban population and male are 3% more than female population.

Indian workers are grouped into Main workers, Marginal workers and Non-Workers. Marginal workers are further grouped into those working less than 3 months a year and 3 to 6 months a year. Within Marginal and Non-workers, people are available for work.

Main workers contribute to about 1/3rd in the entire population.

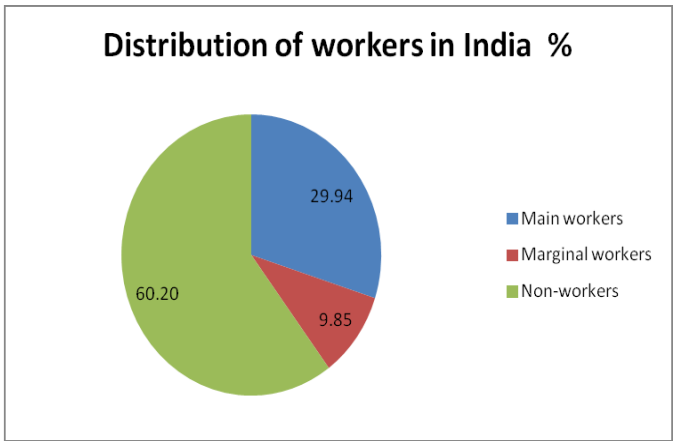


Figure II. Distribution of workers in India

Main workers data is analyzed to know the distribution across male, female, rural and urban areas.

In all the age groups, male contribute to 3/4th of total main workers and female are only 1/4th. Main workers in urban areas are almost close to main workers in the rural India up to the retirement age of 60 years. Late rural main workers are than urban main workers.

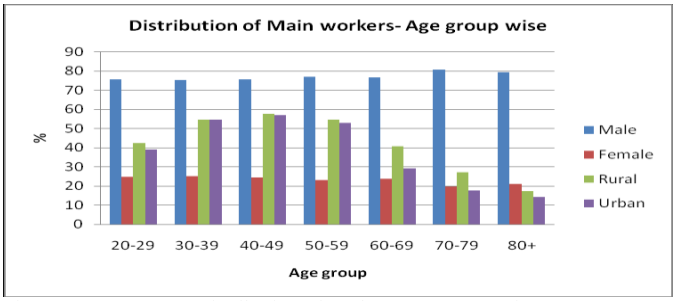


Figure III. Distribution of workers-Age group wise

Female Non-workers are about 90% from 30 to 60 years. Male non-workers are significant after the age of retirement. We can also find significant male non-workers in the age group of 20-29.

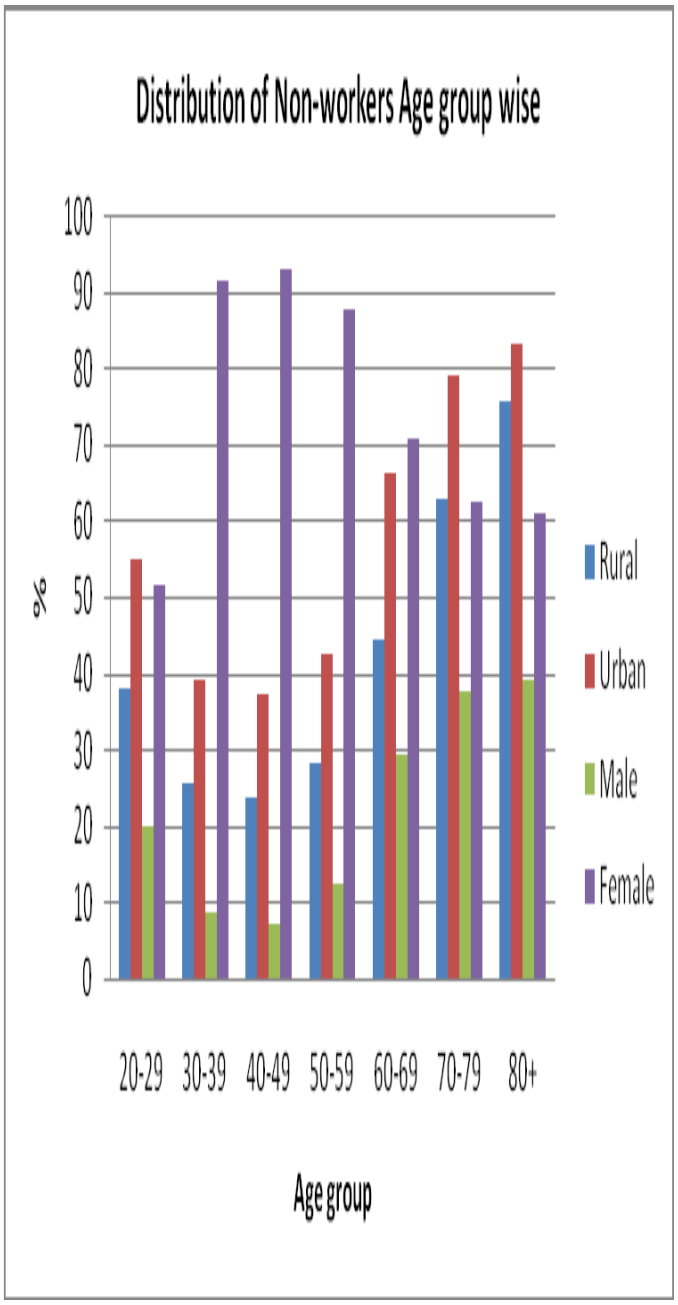


Figure IV. Distribution of non workers age group wise

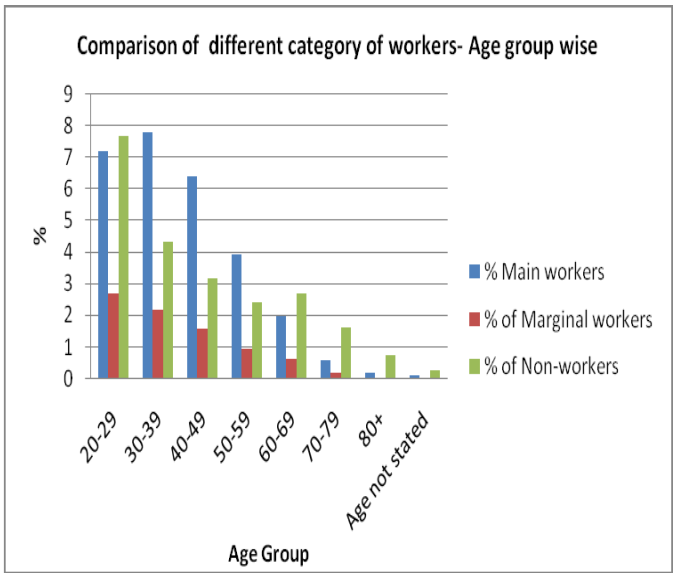


Figure V. Comparisonof different category of workers-Age group wise

From the above Figure , We can see more Non-workers than Main workers in the age group of 20-29. Later, we can find significant domination of main workers compared to Non-workers till age of 60. Good number of main workers can also be found in the age group 60-69

IV CONCLUSION

Big data technology requires predefined formats for data representation and can Process structured, semi-structured and un-structured data. With reduction in storage costs, organizations can store data for long periods and analyze it to have insights into data. These large data sets can contain valuable information which can be helpful for other agencies, for their planning and policy making. This paper gives the distribution of different category of workers across male-female population and also between Rural-Urban areas

References

[1] https://www.meridium.com/challenges/big-data0
[2] Alvaro A. Cardenas, Pratyusa K. Mandate, Sreeranga P. Rajan, "Big Data Analytics for Security", IEEE computer and- Reliability Societies, 2013, pp. 74-76.
[3] Jeff Markey, " How to Manage Big Data's Big Security Challenges"
[4] "Cloud Security Alliance Top Ten Big Data Security and Privacy Challenges", Cloud Security Alliance, 2012.

[5] "Cloud Security Alliance Big Data Analytics for Security Intelligence", Cloud Security Alliance, 2013.
[6] Jon Oltsik, " The Big Data Security Analytics Era Is Here", Enterprise
[7] Strategy group, 2013
[8] http://www.brickmarketing.com/define-log-file.htm
[9] https://downloads.cloudsecurityalliance.org/ initiatives/bdwg/Big_Data_Analytics_for_ Security_Intelligance.pdf
[10] http://www.censusindia.gov.in/2011census/B-series/B-Series-01.html
[11] Ian Robison, Jim Webber and Emil Eifrem. *Graph* Database. CA : O'Reilly Publishers

An Intelligent Prediction of accidents for Domotics

S Jahnavi
jahnavi_s@vnrvjiet.in
Assistant Professor, CSE
VNR VJIET

Deepa Kumari
deepasneha10@gmail.com

***Abstract*— Domotics are not only integration of technology and control of appliances in the home but also provide services for a better quality of life. But if we want some extended facility for the handicapped and old aged people to alert them in the critical situations in quick response way. All these things are feasible in the smart home model but for these intimated technological system to work we require some sort of tiny end point monitoring device called Smart dust motes that will feed back basic information what is happening in the home or in the air. This paper describes that how Smart motes are considerably proved to be a masterpiece that can fit into anyplace and anywhere so as to detect even small vibrations and can predict accidents caused by malfunctioning of the devices and there by alerting the people**

Keywords- Domotics, Smart Dust, Smart motes, Automation, Home appliances

I. INTRODUCTION

Home Automation is where daily things are being completed automatically without any human interference. Many basic tasks such as turning on or off of certain devices can be controlled remotely. When the control these devices is beyond human control, the process of monitoring and reporting them back becomes a key and predominant feature. Automation being an industrial application migrated to homes as a result of rapid growth in development of low cost consumer electronic devices. Home automation is the concept of controlling all home appliances connected to a central location, either remotely or with in a close proximity.

With the arise of wide range of electronic components in the market it becomes necessary to connect all of them to one central component for monitoring and controlling them easily. From many years there have been many home automation systems being developed to provide a networked control of the devices [1]. Cellular mobile technology emerged as a powerful tool integrating the mobile technology and home automation, there by providing a mobile phone-based home automation [2]. With a rapid growth of the internet control of these electronic devices from a remote location becomes a potential need. However, with an increase in variety of home appliances the connectivity of these devices still remains unexploited.

"Domotics" also referred as Home automation is formed by combining a Latin word "domus" which means home and the words informatics, telematics and robotics. With the decrease in the size of computing devices increased the small computing elements which increased the opportunities for ubiquitous researchers to reshape the interaction between people and computers.

"Smart dust" an emerging technology made of many tiny microelectromechanical systems (MEMS) is used to sense and communicate with the surroundings. The ideology behind this system is to combine a variety of sensors devices that detect light, temperature and many more, tiny computers and wireless communicators in to a small cubic-millimeter mote and develop a distributed sensor network. This paper proposes a smart home by integrating dust motes technology which helps the residents to detect and predict the unusual action of home appliances there by reducing the risks of accidents.

II. EXISTING HOME BASED AUTOMATION

A lot of research has done in the field of home automation and many definitions were defined by different authors. It was in late 1970's when the concept of home automation came around. Automation reduces wastage and makes an efficient utilization of electricity and water [13].

Ali Ziya and Umit Buhur [14] proposed a low cost and flexible internet based home automation system which can control multifunctional devices using wireless communication. Alheraish [1] designed a M2M wireless system which transmits and receives data remotely over a network. M2M also known as machine-to-machine, man-to-machine or mobile-to-machine proved to be cost efficient. The author implemented this system using a various communication networks out of which cellular networks gained much attention.

Khusvinder et al. [3] presented another low cost and flexible ZigBee based home automation system which connects home appliances designed by multiple vendors. This system allows all the devices connected to the system to be monitored and controlled through a variety of control devices. ZigBee based remote control was one of these devices including others as any Java supported wifi devices.

Han et al. [10] proposed a user friendly home automation system based on 3D virtual world. The author tried to provide a realistic user interface through which a user can control and

monitor the devices anywhere through internet. Amaya Arcelus [15] and S. Nourizadeh [16] proposed systems which focused more on providing different health services through the smart home system. Rialle [6] later discussed on the challenges and innovations in smart health homes.

A. What is a Home automation system made up of?

Many forms of hardware and communication interfaces are available which are used in designing a smart home. Some of them include usage of additional communication and control wiring, some embed signals in the existing power circuit of the house, some use radio frequency (RF) signals, and some use a combination of several methods. Figure 1 [17] shows hardware layout for a home automation system.

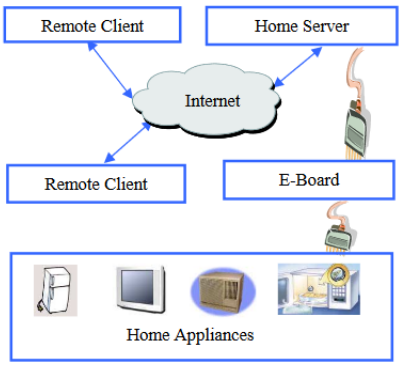


Figure 1: System hardware layout.

The elements of a domotics system are: hardware controllers or software controllers, sensors, actuators and an interface through all these devices can be controlled. A centralized controller can be used, or multiple intelligent devices can be distributed around the home.

Interconnection can be a wired, wireless or an infrared connection. For a wired connection we can use an optical fiber, coaxial or twisted pair cable and a powerline. WI-FI, GPRS and UMTS, Bluetooth, ZigBee, or a Z-wave can be used for wireless interconnection setup. If the interconnection is by infrared Consumer IR is used.

We have a different home automation standards based on different control systems [19], some of them are KNX, ZigBee, Z-wave, EHS(European Home Systems), X10, LonWorks, INSTEON

B. What is smart dust?

Smart dust was developed by Professor Kristoder Pister(UC Berkeley) in the year 1998. The idea was to develop a Micro Electro-Mechanical System (MEMS), which is a package of small sensors, tiny computers that can monitor and compute tons of data. Smart Dust facilitates applications like surveillance to defence related services, inventory control, product quality monitoring, smart offices and many more [18]. The dust mote contains a microcontroller which performs the assigned tasks and also controls the energy provided to the other components of the system. The sensors process the data captured from the surrounding periodically and sends the readings to the microcontroller where this data is stored. It transmits the data through available communication networks. The main constraint in design of these tiny motes is energy consumption by different sensors and devices. Timers are maintained for efficient usage of energy. Most of the devices are powered up only by receiving signals from these timers, where the other time they are powered off. The magnified view of a smart dust mote is shown in figure 2.

Figure 2: Smart Dust Mote

The Key Components of Smart Dust

- A thick-film battery, a solar cell with a charge-integrating capacitor or combination of both
- An optical photo diode for data detection and receiving
- Different small sensors from sensing various types of data
- A semiconductor laser diode and MEMS beam steering mirror for active optical transmission
- Corner Cube Retro-reflector(CCR) for passive optical transmission
- A signal processing and control circuitry

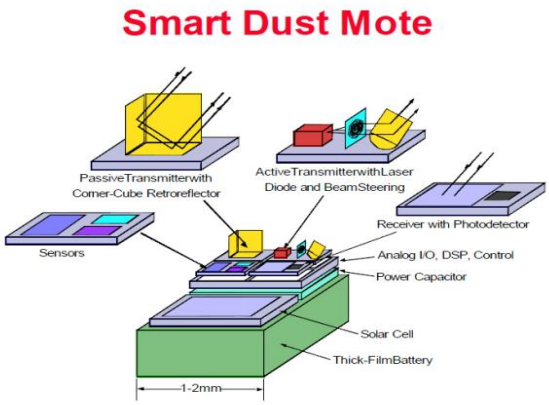


Figure 3: Major Components of Smart Dust mote

Identify applicable sponsor/s here. *(sponsors)*

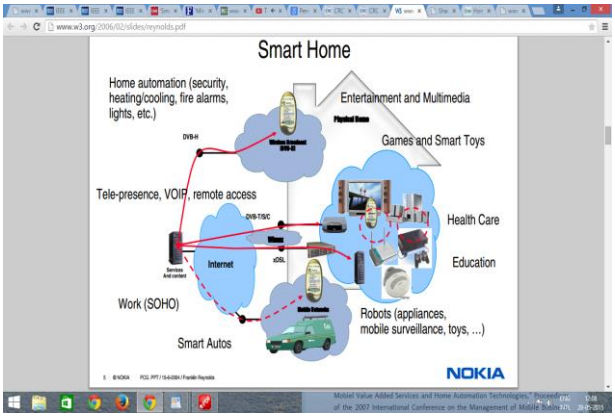
Smart dust is visualized to be part of many application domains such as environmental monitoring, military systems, etc. Due to its tiny size, smart dust is expected to enable a number of novel applications. Smart dust is also used in tracking the location of real world phenomena [7].

C. *Analysis of the Existing Home Automation Systems*

Although the advancement of home automation system is at unprecedented rate there are some problems limiting the consumers to adopt the technology. The present smart homes available are very expensive and affordable by only upper class families [2] [5] [11].

The present-day systems installation is very difficult and complex. Professionals are required to install and configure the electronic devices and their control systems. It is seen that the firms are focusing on either the software or the hardware but not both. This results in poor interfaces between the components and the control system technology [3]. The current systems tend to provide services for any one device category but not all. Thus, to control variety of electronic appliances the customers must correlate among different devices and their brands [3] [4].

Also there are few home automation systems which provide both monitoring and controlling of the system where most of the systems provide only one of the functionality. If anyone wants to remotely control the home and also constantly monitor the power usage of devices, they must provide a way to interface two different systems from two different companies [3]. Also the Current systems are not very adaptable by the end user. This is restraining the users to do different from what the system is intended to do and also doesn't take account of the users daily needs.



Smart Home

III. PROPOSED SYSTEM DESIGN

We propose the design of a security system in home automation by merging the application of the smart dust motes. The main purpose is to help handicapped and old aged people which will enable them to control home appliances and alert them in critical situation tasks. Smart Dust describes about microelectromechanical (MEMS) devices that include robust sensors, computational ability and communication subsystem. The Machine-to-Machine accessible smart dust is made of motes, which are tiny sensors capable of spreading throughout buildings and into the atmosphere to collect and monitor data [20].All of the motes in the area create a giant, amorphous network that can collect data. Data funnels through the network and arrives at a collection node, which has a powerful radio able to transmit a signal many miles. When some fault and error occur, the motes that detect it transmit their location and their sensor readings. Neighboring motes pick up the transmissions and forward them to their neighbors and so on, until the signals arrive at the collection node and are transmitted to the destined node. The node can now display the data on a screen and see, in real time, the point where fault is present through the field of motes. Some allow sensors to move into places where they have not been before and others reduce the time needed to read sensors individually.

IV. IMPLEMENTATION

Smart Dust motes are run by microcontrollers. These microcontrollers consist of tiny sensors for recording various types of data. Timers are used to run these sensors. These sensors do the job of collecting the data. The data obtained are stored in its memory for further interpretations. It can also be sent to the base controlling stations.

CCR that comprises of three mutually perpendicular mirrors of gold coated poly silicon has the property that any incident ray of light is reflected back to the source provided that is incident within a certain range of angles centered about the cubes body diagonal. The micro fabricated CCR includes an electrostatic actuator that can deflect one of the mirrors at kilohertz rate. Hence the external light source can be transmitted back in the form of modulated signal at kilobits per second.

Although a passive transmitter can be made more omni directional by employing several CCR"s oriented in different directions, at the expense of increased dust mote size[8].

Uses

- Detection of corrosion in aging pipes before they leak.
- Monitoring of humidity, temperature, vibrations, dust, aeration.

Glue a dust mote on each your fingernails. Accelerometers will sense the orientation and motion of each of your fingertips, and talk to the computer in your watch. QWERTY

is the first step to prove the concept, but you can imagine much more useful and creative ways to interface to your computer if it knows where your fingers are: sculpt 3D shapes in virtual clay, play the piano, gesture in sign language and have to computer translate. Combined with a MEMS augmented reality heads-up display, entire computer I/O would be invisible to the people around you. Couple that with wireless access and you need never be bored in a meeting again. Surf the web while the boss rambles on and on.

A. Inventory Control

The carton talks to the box, the box talks to the palette, the palette talks to the truck and the truck talks to the warehouse, and the truck and the warehouse talk to the internet. Know where your products are and what shape they are in anytime, anywhere. Sort of like FedEx tracking on steroids for all products in your production stream from raw materials to deliver goods.

Interfaces for disabled: put motes on a quadriplegic's face, to monitor blinking & facial twitches – and send them as commands to a wheelchair /computer/other device. This could be generalized to a whole family of interfaces for the disabled. Your house and office will be aware of your presence and even orientation in a given room. Lighting, heating and other comforts will be adjusted accordingly

V. CONCLUSION AND FUTURE ASPECTS

According to researchers, the smart dust will play a very beneficial role of monitoring every element of our earth. It is likely to be the core technology of the future world. The use of motes in home automation techniques advances the level of technology in the most consistent way to track widespread automated syetem. The battery-powered matchbox-sized Motes can be built to sense numerous factors, from light and temperature for energy saving applications to location to dynamic response, ultimately for domotics purpose so that it can intimate old aged or handicapped people and alert them for any critical situations.

With the recent development in the field of the network of structural Motes, it also extends the researchers to use computer simulations of earthquakes, fires, or other structural threats to forecast the potential for damage. We must say there will be disaster risk reduction, expecting the flood of data from the Smart Dust Motes to greatly increase the accuracy of finite element analyses, a method of computer modeling where mathematical equations represent a structure's behavior under certain conditions. We must say, Smart dust researchers are helping to monitor the world so that in future it will benefit people as well as the environment.

REFERENCES

1. Alheraish, A. "Design and implementation of home automation system." *IEEE Transactions on Consumer Electronics* 50.4 (2004): 1087-1092.
2. Andreas Rosendahl and Goetz Botterweck, "Mobile Home Automation Merging Mobiel Value Added Services and Home Automation Technologies," Proceedings of the 2007 International Conference on the Management of Mobile Business, Toronto, Canada, 9-11 July 2007.
3. Gill, Khusvinder, et al. "A zigbee-based home automation system." *IEEE Transactions on Consumer Electronics* 55.2 (2009): 422-430.
4. W. Keith Edwards, Rebecca E. Grinter, Ratul Mahajan, David Wetherall (2011) Advancing the State of Home Networking. Communications of the ACM Vol. 56, No. 6.
5. A.J. Bernheim Brush, Bongshin Lee, Ratul Mahajan, Sharad Agarwal, Stefan Saroiu, Colin Dixon (2011) Home Automation in the Wild: Challenges and Opportunities. CHI 2011.
6. Rialle, Vincent, et al. "Health" smart" home: information technology for patients at home." Telemedicine Journal and E-Health 8.4 (2002): 395-409.
7. M. Weiser and J.S. Brown, "The Coming Age of Calm Technology," http://www.ubiq.com/hypertext/weiser/acmfuture2endnote.htm, 1996.
8. Römer, Kay. "Tracking real-world phenomena with smart dust." Wireless Sensor Networks. Springer Berlin Heidelberg, 2004. 28-43.
9. Kahn, Joseph M., Randy H. Katz, and Kristofer SJ Pister. "Next century challenges: mobile networking for "Smart Dust"." Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking. ACM, 1999.
10. Han, Jinsoo, et al. "User-friendly home automation based on 3D virtual world." *IEEE Transactions on Consumer Electronics* 56.3 (2010): 1843-1847.
11. http://catchupdates.com/smart-dust/#sthash.8fypPzwq.dpuf
12. http://seminarprojects.com/Thread-smart-dust-full-report#ixzz3dISxOIZI
13. Shih-Pang Tseng, Bo-Rong Li, Jun-Long Pan, and Chia - Ju Lin, "An Application of Internet of Things with Motion Sensing on Smart House", 978-1-4799-62846/14 © 2014 IEEE.
14. Alkar, Ali Ziya, and Umit Buhur. "An Internet based wireless home automation system for multifunctional devices." *IEEE Transactions on Consumer Electronics* 51.4 (2005): 1169-1174.
15. Arcelus, Amaya, et al. "Integration of smart home technologies in a health monitoring system for the elderly." *Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on*. Vol. 2. IEEE, 2007.
16. Nourizadeh, Shahram, et al. "Medical and home automation sensor networks for senior citizens telehomecare." *2009 IEEE International Conference on Communications Workshops*. IEEE, 2009.
17. Al-Ali, Abdul-Rahman, and Mohammad Al-Rousan. "Java-based home automation system." *IEEE Transactions on Consumer Electronics* 50.2 (2004): 498-504.
18. Puccinelli, Daniele, and Martin Haenggi. "Wireless sensor networks: applications and challenges of ubiquitous sensing." *IEEE Circuits and systems magazine* 5.3 (2005): 19-31.
19. http://www.slideshare.net/olafusimichael/500project1

20. https://chaione.com/blog/smart-dust-communication-systems-and-the-future-world/

Introducing Cloud in Remote Sensing and Instance Creation using OpenStack

G Maneesha[#1] , K Praveen Kumar[#] , M Manju Sarma[*] ,V Manikumar[*]

[#]*Department of Computer Science, VRSEC,*
Vijayawada, Andhra Pradesh, India.
[1] mani90gudapati@gmail.com
[*]*Software Group, National Remote Sensing Center – ISRO,*
Bala Nagar, Hyderabad, Telangana, India.

Abstract **— Remote sensing enables to observe the earth surface from the far away distance by imaging. To manage those images backend infrastructure needed that is provided by Cloud. Cloud computing provides the pool of resources for efficient storage and processing. This paper gives the basic introduction to remote sensing and cloud computing. This paper also listed the various models and characteristics of cloud. With the help of the open source tool like OpenStack, how an instance can be created was discussed.**

Keywords. Cloud computing, Remote Sensing, OpenStack, Instance.

I. INTRODUCTION

Remote sensing is a process of observing, recording and perceiving objects on the earth's surface. In this process(Fig. 1.), the sensors are not in direct contact with the objects that are being observed. Through an intervening medium, physical carrier passes the information from the objects to sensors. In remote sensing usually electromagnetic radiation is used as an information carrier. The observed scenes are taken in the form of images, which are the outputs of remote sensing system. These images are generally in the form of digital images. To extract the useful information from the image, image processing techniques like image enhancement, analysis and interpretation may be employed, which helps in enhancing, correcting or restoring the image if the image has been subjected to geometric distortion, blurring or degradation by other factors[1]. To speedup the processing of the images send by the Geo-orbit Satellite (sends images 24*7) and to handle the huge number of these images, we are introducing the cloud concept at the processing level of the Remote Sensing.

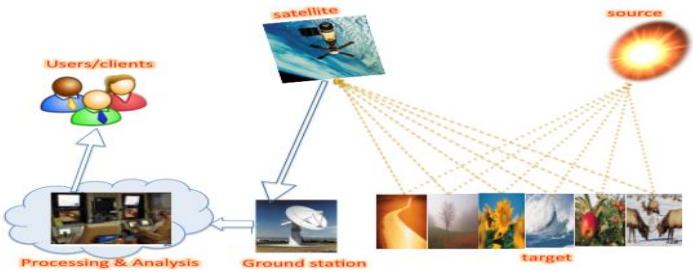


Fig. 1. Remote Sensing Process

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (The NIST Definition of Cloud Computing) [2]. Simply, cloud is an advanced version of virtualization. Virtualization is the concept of creating virtual machines that can be used as an operating system, server, network resources or storage devices.

Cloud computing (Fig .2.) explained using cloud computing stack, which manages the hardware/software of data center into respective service layers.

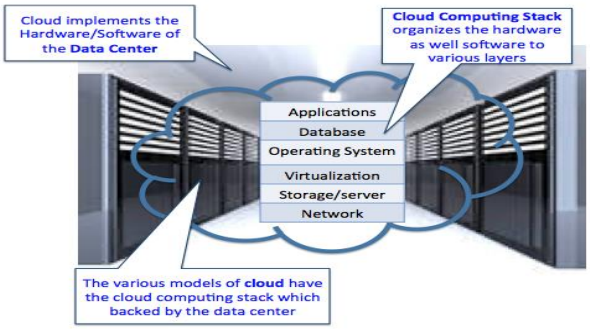


Fig .2. Cloud Computing Stack

A. *Characteristics of Cloud Computing*

Based on the NIST Definition of Cloud Computing, there are five essential characteristics. They are:

- *On-demand self service*: Cloud facilitates the on-demand services like compute, storage to the users without any human interaction.
- *Broad network access*: Network provides the capability to access the heterogeneous platforms (thin and thick clients, mobiles, laptops)
- *Resource pooling*: The pool of resources that are location independent are dynamically served to the users.
- *Rapid elasticity*: Cloud enables the service-provider to scale-in and scale-out their resources without affecting user applications.
- *Measured service*: Cloud provides the metering services, which helps the users to pay for what they use.

B. *Cloud Computing Architecture*

The architecture of cloud consists of many loosely coupled components. Fig. 3. depicts cloud computing architecture can be broadly divided into two parts, front end and back end. Network provides the connection between the two ends [3].

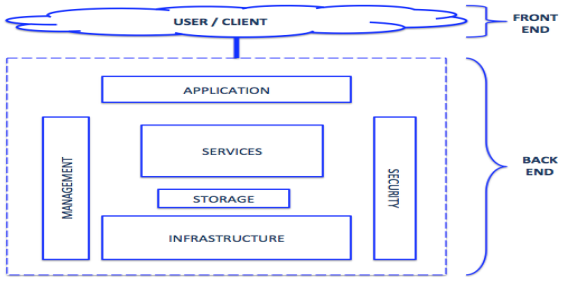


Fig. 3. Cloud Computing Architecture.

The front end of the cloud architecture is the client part that contains the interfaces and applications to access the cloud-computing platform. A web browser is one of the example for this. The back end of the cloud architecture is cloud itself, which contains all resources. These resources provide the cloud computing services such as virtual machines, servers, storage, etc.

C. *Models of cloud computing*

Cloud computing models are categorized into two types namely deployment models and service models [4].

1. *Deployment model*:

Based on the size, access and ownership the deployment models are classified into four types, which shows in Fig. 4. They are:

- *Private cloud* is implemented with organization's own network connection.
- *Public-cloud* is provided by service provider and shared among organizations or by the standalone user depending on their requirements and pay for the services they used.
- *Hybrid-cloud* is the combination of private cloud and public cloud or any number of clouds.
- *Community-cloud* is the variant of any cloud with group of people having common interest.

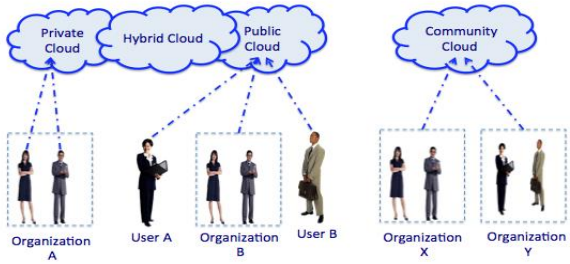


Fig. 4. Types of Deployment Model

2. *Service model:*

Based on the service provided to the end user, the service models are classified into three types that can be seen in Fig.5. They are:

- *Infrastructure as a Service (IaaS)* provides physical components like networks, virtual machines, and storage to clients.
- *Platform as a Service (PaaS)* provides the run-time environment for developing and deploying the applications.
- *Software as a Service (SaaS)* provides the specific application according to the user's requirement.

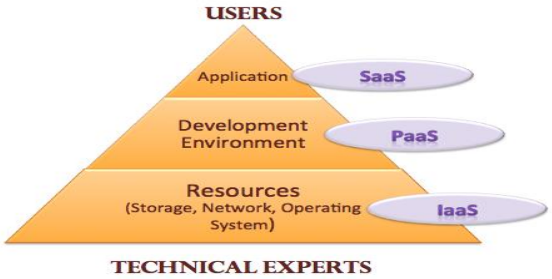


Fig. 5. Types of Service Types

II. LITERATURE SURVEY

Shefali Aggarwal 2004 explained about how the remote sensing technique works and importance of remote sensing imagery in different applications. Basic principles of remotely sensed data collection mechanism were discussed [1]. Timothy Grance and Peter Mell 2011 from NIST proposed the definition of the cloud computing and listed the five essential characteristics, four deployment models and three service models [2].

Vikas Goyal. 2012 discusses how cloud computing architecture has provided the services through its layers and their development models [3]. Rahul Bhoyar, Nitin Chopde 2013 explores about different service models (IaaS, PaaS, and SaaS), cloud computing types (public, private, community) and advantage and usage of cloud. They also discuss the cloud database including deployment model, characteristics and identifies some technological and legal issues in cloud computing [4].

Girish L S1, Dr. H S Guruprasad 2014 discuss about implementing a private cloud using open source software and operating system. OpenStack is open source software used by the developers to run the cloud. OpenStack services can be accessed through API's. The important components of OpenStack are Nova, Keystone and Glance, Keystone and Horizon [5]. So to establish the IaaS, open source tool OpenStack is used [6].

III. INFRASTRUCTURE AS A SERVICE USING OPENSTACK

One of the open-source tools used to establish the private cloud is OPENSTACK. OpenStack is a set of software tools for establishing and maintaining cloud-computing platforms for both public and private clouds. OpenStack began in 2010 as a joint project of Rackspace and NASA. Companies like IBM, HP etc. are using OpenStack as a base for their cloud deployments. Apart from installing, configuration is the major job that involves configuring software and synchronizing with hardware to provide virtual cloud services.

OpenStack provides diverse services by its components that are listed in [TABLE I.]

TABLE I
OPENSTACK COMPONENTS

Component	**Service Provided**
Keystone	Identity service (authenticate and authorize the users, projects, roles and other services).
Nova	Compute service (heart of the OpenStack which manages the instances)
Glance	Image service (stores the virtual image templates)
Horizon	Dashboard service (provides the GUI)
Neutron	Networking service (Manages the instance network)
Swift	Object Storage service (For storage)
Cinder	Block Storage service (Manages the volumes)
Trove	Database service (used in creating the database instance)
Heat	Orchestration service (helps to deploy the applications using .yaml templates)
Ceilometer	Metering service (used for billing the resource usage)

IV. INSTANCE CREATION

In cloud, instance works like a virtual machine. Instance can be operating system, volume or database. Instance created with the help of flavor and an image. Flavor is a virtual hardware template and image is a virtual software template. Before launching an instance the following environment need to be setup:

a) Check whether the available flavors match the instance requirement.
   - List available flavors: OpenStack flavor list / nova flavor-list
   - If the available flavors do not match the

instance requirement then a new flavor can be created using the dashboard (GUI) or command line.

nova flavor-create [name id vcpus disk ram]

b) Image also can be created using dashboard or command-line.

nova image-create[name image-location format]

c) Network should be established by providing the network-name, IP address ranges, gateway, allocation-pool, sub-net.

d) Key-generation for secure login to the instance using key-gen command.

e) Creating security-groups to enable tcp, udp, icmp and other protocols to access the instance.

f) Launching an instance.

nova boot [flavor-name image-name network-id key security-group] instance-name

g) After launching an instance the following process (Fig.6.) will be performed where the instance transforms from build state to active state.

(1) Dashboard sends the REST call to keystone for authenticating the user credentials. Keystone sends back auth-token for communicating with other components using REST-call.

(2) The newly launched instance request is converted to REST API request by dashboard/CLI.

(3) REST-API request is send to nova-api, where it validates the token with the keystone which sends back the updated auth header with roles and permissions.

(4) Nova-api interacts with nova-database for creating the new instance entry.

(5) The rpc.call request sent by nova-api to nova-scheduler is picked up from the queue and then scheduler interacts with the nova-database to find the appropriate host that is updated with the instance entry.

(6) nova-compute picks the rpc.cast request from the queue sent by the nova-scheduler and to get the information of an instance it sends the rpc.call request to the nova-conductor.

(7) After picking the request from the queue the nova-conductor interacts with the nova-database and returns the instance information to the nova-compute.

(8) nova-compute gets the image meta-data by validating the auth-token sent to glance-api through keystone.

(9) Nova-compute acquires the instance IP by sending the auth-token to neutron-api. With the help of keystone, neutron server validates the auth-token and sends back the network-information to nova-compute.

(10) Using libvirt or api nova-compute loads the information on to the hypervisor driver and spawns the request.

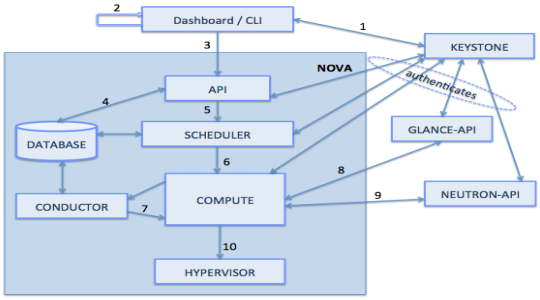


Fig.6. Flow of Instance Creation

h) After instance state became active, it can be accessed through vnc console or by ssh.

i) Image processing application can be deployed on the instance by installing the required rpms.

j) These instances will be served to the authorized users.

V. CONCLUSION

Cloud computing is a powerful technology to handle the huge amount of data or service. This paper mainly focused on basics of cloud computing and presented how to create the instance using the OpenStack tool (IaaS). Each step while creating the instance was also discussed. The work can be further extended by creating the database instance, providing volumes to the instance or creating the dynamic instance.

REFERENCES

[1] Shefali Aggarwal, "Principles of Remote Sensing. Satellite Remote Sensing and GIS Applications in Agricultural Meteorology", pp.23-38, 2004.

[2] Timothy Grance and Peter Mell, "The NIST Definition of Cloud Computing", *NIST Special Publication* 800-145, Sept. 2011.

[3] Vikas Goyal, "Layered architecture of Cloud Computing", *International Journal of Computing & Business Research*, ISSN 2229-6166 (2012).

[4] Rahul Bhoyar, Nitin Chopde, "Cloud Computing: Service models, Types, Database and issues", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol 3, issue 3, 2013.

[5] Girish L S1, Dr. H S Guruprasad, "Building Private Cloud using OpenStack", *International Journal of Emerging Trends & Technology in Computer Science*, ISSN 2278-6856 Volume 3, Issue 3, 2014.

[6] Aniruddha S. Rumale, D.N.Chaudhari, "Cloud Computing: Infrastructure as a Service", *International Journal of Inventive Engineering and Sciences* ISSN: 2319–9598, vol-1, issue-3, 2013.

Aggregate Cryptosystem based Data Sharing in Distributed Computing

K Nagendra[1], K Leela Prasanth[2], K Praveen Kumar[3], K Venkateswara Rao[4]
[1]M.Tech, Department of Computer Science, VR Siddhartha Engineering College, Kanuru, Vijayawada, AP, India
[2]M.Tech, Department of Computer Science, VR Siddhartha Engineering College, Kanuru, Vijayawada, AP, India
[3]Sr.Ass.Professor, Department of Computer Science, VR Siddhartha Engineering College, Kanuru, Vijayawada, AP, India
[4]Research Scholar, MGR University, Maduravoyal, Chennai, India

Abstract: Cloud computing has showed up as one of the most significant paradigms in the IT market lately. Since this new handling technology needs clients to believe in their useful information to reasoning providers, there have been enhancing security and comfort problems on shortened details. A few strategies using quality based security (ABE) have been prescribed for access administration of abbreviated subtle elements in thinking processing; be that as it may, the greater part of them experience from resoluteness in actualizing complex accessibility administration rules. The one of a kind is that one can add up to any arrangement of key imperative variables and make them as conservative as a solitary key, however covering the force of the considerable number of keys being accumulated. By considering these issues in real time secure cloud data sharing, in this paper we propose to develop Key aggregate cryptosystem with real time data stream management. In different terms, the key proprietor can discharge a consistent size aggregate key for flexible alternatives of figure composed content set in distributed storage space, however the other secured data documents outside the set stay classified. This lightweight aggregate key can be in a perfect world sent to others or be spared in a smart card with extremely confined ensured storage room. We give official security examination of our procedures in the ordinary configuration. We likewise clarify other project of our systems.

Index Terms: Cloud Computing, Attribute based encryption, Scalable and reliable data encryption and decryption, secure Hashing.

I. INTROUDCTION

Distributed computing is a model for empowering pervasive system access to share the configurable PC assets. Distributed computing and stockpiling choices furnish clients and organizations with different capacities to store and procedure their data in outsider data offices [1]. It relies on upon talking about of sources to accomplish reasonability and monetary frameworks of extent, like an application (like the force network) over a framework. At the base of cloud preparing is the more extensive thought of consolidated offices and disseminated administrations.

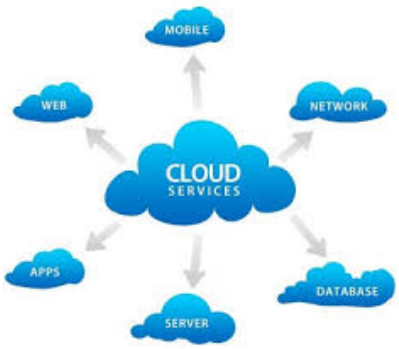


Figure 1: Cloud computing services in resource monitoring.

As shown in the above figure cloud computing provides three types of solutions regarding thinking support and other proceedings present in distributed handling functions. SAAS(Software As a Service), PAAS(Platform As a Service), and Facilities As a Service are three solutions of the thinking handling for storage space information, handling information and preserves of

information which includes all the activities of the customers presentation may appears recent development of information motivation program [2]. Consider the examples of Mediafire.com, SendSpace.com and Amazon Cloud Web solutions and other solutions are storage space of information in thinking and other continuing website signing up procedure. These are the successive web sites for providing solutions to various customers for storing their information with handling program. Reasoning contains share of solutions of details. All kinds of customer demands are applied with good performance and interaction expense contains high. Protection and comfort signify major issues in the adopting of reasoning technological innovation for information storage. A strategy to minimize these issues is the use of security. Be that as it may, though security ensures the protection of the data against the thinking, the utilization of ordinary security procedures is not adequate to bolster the organization of fine-grained business availability control Policies (ACPs). Numerous organizations have today ACPs controlling which clients can openness which information; these ACPs are frequently demonstrated as far as the characteristics of the clients, for the most part known as distinguishing proof components, utilizing availability administration dialects, for example, XACML. Such a methodology, for the most part known as property based availability controllability (ABAC), encourages fine-grained openness administration which is pivotal for high-affirmation data security and solace.

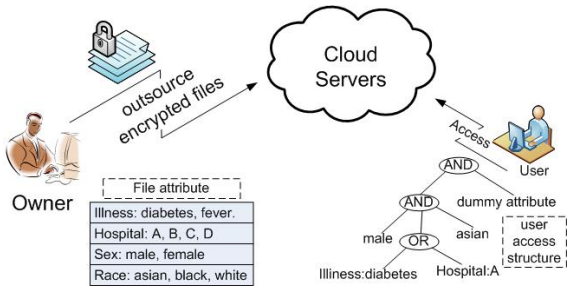


Figure 2: Attribute based encryption for outsourcing data [2].

Attribute-Based Encryption (ABE) allows only organizations having a specified set of features can decrypt cipher texts [3][4]. ABE is appropriate to accessibility management such as the computer file discussing techniques, because several organizations can be provided for the decryption of a cipher text. We have been suggesting an enhanced ABE plan that is more effective than past one. Through present delegate calculations we are going to consume the solutions usage with new security difficulties execution procedure. In the storage space service program, the reasoning can let the customer, information proprietor to shop his information, and discuss this information with other customers via the reasoning, because the reasoning can provide the pay as you go atmosphere where people just need to pay the money for the storage space they use. For defending the privacy of the saved information, the information must be secured before posting to the reasoning. The security plan used is attribute-based security. The ABE plan used a customer's identification as features, and a set of features were used to secure and decrypt information. One of the main efficiency disadvantages of the most current ABE techniques is that decryption is costly for resource-limited gadgets due to coupling functions, and the number of coupling functions required to decrypt a cipher written text develops with the complexness of the accessibility plan. The ABE plan can outcome the issue that information proprietor needs to use every approved customer's community key to secure information.

Trust that Alice puts all her own pictures on Drop Box, and she wouldn't like to uncover her pictures to everybody. Because of different data spill likelihood Alice can't experience treated by simply relying upon the solace insurance components offered by Drop Box, so she encodes every one of the pictures utilizing her own imperative variables before posting. One day, Alice's mate, Bob, asks for her to talk about the pictures assumed control over every one of these decades which Bob appeared in. Alice can then utilize the examiner work of Drop Box, yet the issue now is the manner by which to allot the unscrambling rights for these pictures to Bob. A conceivable decision Alice can pick is to securely convey Bob the key vital variables locked in. Normally, there are two intemperate systems for her under the conventional security world view.:

 Alice scrambles all information records with one and only security key and gives Bob the relating key straight.

Alice scrambles information records with exceptional imperative components and conveys Bob the relating key critical variables.

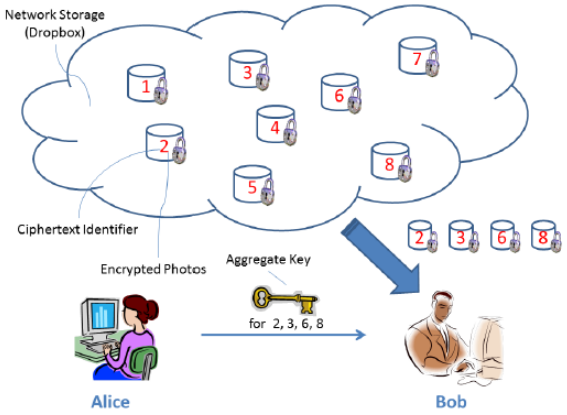


Figure 3: Alice stocks information with identifiers 2, 3, 6 and 8 with Bob by delivering him only one total key.

As demonstrated in figure 3, clearly, the first strategy is lacking since all unchosen information may be likewise discharged to Bob. For the second technique, there are sensible issues on execution. The quantity of such imperative elements is the same number of as the mixture of the common pictures, say, a million. Moving these mystery keys ordinarily needs a secured course, and putting away these essential elements needs rather extravagant ensured storage room [6]. The expenses and muddlings connected with typically enhance with the mixed bag of the unscrambling essential variables to be conveyed. In a nutshell, it is exceptionally gigantic and extravagant to do that. Encryption essential components likewise accompany two tastes — symmetric key or lopsided (open) key. Utilizing symmetric encryption, when Alice needs the data to be begun from a third festival, she needs to give the encryptor her mystery key; clearly, this is not generally suitable. By complexity, the security key and decoding key are distinctive in broad daylight key security. The utilization of open key encryption gives more adaptability for our projects. For instance, in business designs, each specialist can transfer encoded data on the thinking storage room server without the data of the organization's expert mystery key.

In this way, the best solution for the above issue is that Alice scrambles information records with novel open keys, yet just conveys Bob stand out (steady size) decoding key [8]. Since the unscrambling key ought to be sent by means of a protected channel and kept key, minimal key measurement is constantly suitable. For instance, we can not suspect gigantic stockpiling for unscrambling imperative elements in the asset requirement gadgets like advanced cells, astute charge cards or Wi-Fi pointer hubs.

Particularly, these key essential variables are typically spared in the carefully designed capacity, which is generally extravagant. The present investigation activities chiefly focus on minimizing the cooperation particulars, (for example, information exchange utilization, rounds of correspondence) like aggregate mar

The remaining of this paper organized as follows: Section II provides overview of the related work presented in previous application procedures, In Section III present Traditional approach with security considerations; Section III describes effective data presentation and construction of the proposed approach. Section IV analyzes the security cloud with flexible and effective computation with real time performance evaluation and implementation. Section V describes concluded process of cloud security process.

II. BACKGROUND APPROACH

Typically notice system specific control feature based security contracted schema was presented Contrary to the design for typical ABE, a KGSP and a DSP are furthermore involved. . KGSP is to perform helped key-issuing computations to decrease AA complete a range program when a lot of customers create requirements on personal key creation and key-update.

. DSP is to complete allocated expensive features to get over the disadvantage that the decryption level in typical ABE needs a lot of unwanted features at U.

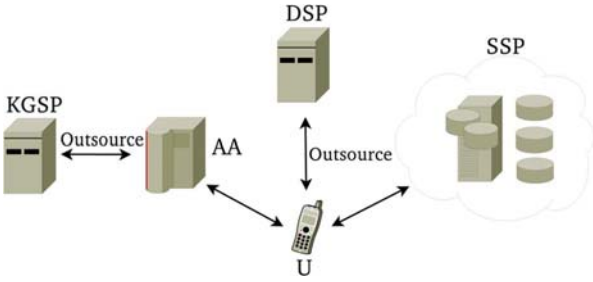


Figure 4: Data outsourcing model using ABE.

Using some of the key demonstration over approximated on the out seeking information with reflection of the secured key with information discussing and other resources using reasoning processing secured solutions such as dedication and other source solutions. We represent (Ienc; Ikey) as the feedback to security and key growth [13, 14] . In CP-ABE plan, (Ienc; Ikey) = (w, A) while that is (w, A) in KPABE, where w and A are function set and availability structure, respectively. Then, depending on the recommended system style, we provide requirements details as follows:

Setup (μ): The set up requirements needs as feedback V a security parameter μ. It results a group key PK and a professional key MK.

Key Gen init (Ikey; MK) : For each user's individual key demand, the initialization requirements for assigned key growth needs as feedback Van availability strategy (or feature set) Ikey and the professional key MK. It results the key several (OKKGSP; OKAA) [2].

Key Gen out (Ikey; OKKGSP): The allocated key creation criteria needs as feedback the availability structure (or function set) Ikey and the key OKKGSP for KGSP. It results a restricted adjustment key TKKGSP.

Key Gen in (Ikey; OKAA): The within key creation criteria needs as feedback the availability structure (or feature set) Ikey and the key OKAA for feature power. It results another restricted modification key TKAA.

Key Sightless (TK): The advance key stunning criteria needs as feedback V the adjustment key TK (TKKGSP; TKAA). It results a individual key SK and a diverted adjustment key f TK.

Secure (μM; Ienc): The security requirements needs as feedback V a idea M and an function set (or accessibility structure) Ienc to be properly secured with. It results the cipher written text CT.

Decrypt out(CT; f TK) : The allocated decryption criteria requires as feedback V a cipher written text CT which was believed to be properly secured under the function set (or accessibility structure) Ienc and the diverted modification key f TK for availability structure (or function set) Ikey. It results the partially decrypted cipher text CT part if (Ikey; Ienc) ¼ 1, otherwise results?, where μ is a predicate pre-specified.

Decrypt (CT part; SK): The unscrambling \requirements requires as info V the somewhat decoded figure content CT part and the individual key SK. It comes about the interesting thought M [2].

Consider the above procedure of secured data outsourcing in thinking may execute viable procedure for security in data proceeding of most recent techniques. Secure outsourcing ABE framework, which helps both secured abbreviated key-issuing and unscrambling. Our new system offloads all accessibility procedure and capacity fitting components in the key-issuing procedure or decoding to a Key Creation Assistance Organization (KGSP) and a Decryption Assistance Organization (DSP), individually, making just a continuous number of straightforward elements for the capacity power and affirmed customers to execute locally. In addition, interestingly, we recommend a contracted ABE development which gives check ability of the abbreviated reckonings results in a viable way. Exhaustive security and execution examination demonstrate that the suggested techniques are checked secured and practical. Powerful Hierarchal structure of the openness control utilizing element based encryption(ABE), better framework was required for amid above concerns effectively..

III. KEY AGGREGATION ENCRYPTION

We first provide the structure and meaning for key total security. Then we explain how to use KAC in a situation of its program in reasoning storage space.

Structure: A key-total security arrangement incorporates five polynomial-time techniques as takes after:

The data proprietor decides the group program parameter through SETUP and produces an open/expert mystery key pair by means of Key Gen. Data can be secured by means of Encrypt by any individual who additionally picks what figure composed content classification is connected with the basically composed content to be secured [8][9] The data proprietor can utilize the expert mystery to produce an aggregate unscrambling key for an arrangement of figure content classes by means of Draw out. The delivered essential variables can be sanction to appoints securely (by means of ensured messages or ensured gadgets) Finally, any client with an aggregate key can decode any figure composed content given that the figure content's classification is contained in the aggregate key by means of Decrypt.

Shared Encrypted Data: Here we explain the primary concept of information discussing in cloud storage space using KAC, shown in figure 3. Suppose Alice wants to discuss her information m1; m2; : : : ;m┤ on the server. She first works Setup (1^{λ}; n) to get param and execute Key Gen to get the public/master-secret key pair (pk; msk). The program parameter param and public-key pk can be published and master-secret key msk should be kept key by Alice. Anyone (including Alice herself) can then protected each mi by Ci = Encrypt (pk; i; mi). The encrypted information are submitted to the server. With param and pk, individuals who work with Alice can upgrade Alice's information on the server. Once Alice is willing to discuss a set S of her information with a buddy Bob, she can estimate the total key KS for Bob by performing Extract(msk; S). Since KS is just a constant size key, it is simple to be sent to Bob via a protected e-mail. After acquiring the total key, Bob can download the information he is approved to accessibility [10]. That is, for each i 2 S, Bob downloading Ci (and some required principles in param) from the server. With the total key KS, Bob can decrypt each Ci by Decrypt(KS; S; i; Ci) for each i 2 S.

IV. IMPLEMENTATION OF KAC

Let G and GT be two cyclic categories of primary purchase p and ^e : □ ×□ $\rightarrow$ □ $_T$ be a map with the following properties:

⦰ Bilinear: $\forall_{g1,g2} \in$ □ $, a, b \in$ □ $, \hat{e}(g\ , g\ ) = \hat{e}(g1, g2)^{ab}$

⦰ Non-degenerate: for some $g \in$ □ $, \hat{e}(g, g) \neq 1$. G is a bilinear team if all the functions engaged above are effectively computable. Many sessions of elliptic shapes function bilinear categories.

4.1. Construction

The style of our primary plan is motivated from the collusion-resistant transmitted security plan suggested. Although their plan facilitates constant-size key important factors, every key only has the energy for decrypting cipher text messages associated to a particular catalog [8]. We thus need to develop a new Draw out criteria and the corresponding Decrypt criteria**.**

Setup: Arbitrarily choose a bilinear team G of primary order p where $2^{\lambda} \leq p \leq 2^{\lambda+1}$ a generator $g \in$ □ $and \alpha \in_R$ □ $_p$. Compute $g_i = g^{\alpha^i} \in$ □ for $i = 1, ..., n, n+2, ..., 2n$. Output parameter as $param = (g, g1, ..., g_n, g_{n+2}, ..., g_{2n})$ Observe that each cipher text category is showed by an index in the integer set $i = 1, ..., n, n+2, ..., 2n$, where n is the maximum variety of cipher text classes.

Key Gen: Pick $\gamma \in_R$ □ $_p$ output the public and master secret key pair: $(pk = v = g^{\gamma}, msk = \lambda)$.

Encrypt: For a message $m \in$ □ $_T$ and an index $i \in \{1, 2, 3, ...n\}$ randomly pick $t \in_R$ □ $_p$ and compute the cipher text $e = (g^t, (vg_i)^t, m.\hat{e}(g1, gm)^t)$.

Decrypt $(K_s, S, i, e = (c1, c2, c3))$**:** If $i \notin S$ output is λ otherwise

$$m = c_3 . \hat{e}(K_s . \prod_{j \in s, j \neq i} g_{n+1-j+i}, c_1) / \hat{e}(\prod_{j \in s} g_{n+1-j}, c_3)$$

4.2. Performance

For protection, the value ^e(g1; gn) can be pre-computed and put in the program parameter. However, we can see that decryption only requires two pairings while only one of them includes the total key [12]. That means we only need one coupling calculations within the protection processor saving the (secret) total key. It is quick to gauge a coupling nowadays, even in asset compelled gadgets. Compelling application usage exist notwithstanding for pointer hubs.

4.3. System Process

The "enchantment" of getting consistent size aggregate key and steady size figure composed content in the meantime originates from the direct size framework parameter.

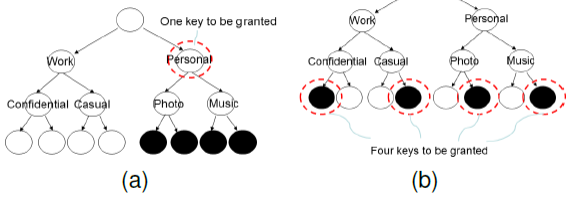


Figure 5: Compact key is not always possible for a fixed hierarchy.

Our motivation is to diminish the ensured storage room and this is an exchange off between two sorts of storage room. The parameter can be put in non-private nearby storage room or in a stockpiling reserve offered by the organization. They can likewise be brought on necessity, as not every one of them are needed in all occasions. The framework parameter can likewise be created by a trusted festival, disseminated between all clients and even hard kept in touch with the client framework (and can be adjusted by means of "patches"). For this situation, while the clients need to trust in the parameter-generator for securely disposing of any transient qualities utilized, the availability control is still guaranteed by a cryptographic mean as opposed to relying upon some server to confine the gets to really.

V. PERFORMANCE EVALUATION

For a substantial assessment, we analyze the territory particulars of the tree-based key undertaking methodology. This is utilized as a part of the Complete Sub tree arrangement, which is a partner solution for the transmitted security issue taking after the surely understood Subset-Cover structure [13]. It uses a set sensible key structure, which is appeared with a complete double key bush of size h (equivalents to 3 in figure 4), and hence can help up to 2h figure composed content sessions, a picked part of which is intended for an affirmed agent.

In an perfect situation as portrayed in Figure 5(a), the delegate can be provided the accessibility 2hs sessions with the possession of only one key, where hs is the dimension a certain sub shrub (e.g., hs = 2 in Figure 5(a)). On the other hand, to decrypt cipher text messages of a set of sessions, sometimes the delegate may have to keep a huge variety of important factors, as portrayed in figure 5(b). Therefore, we are interested in na, the variety of symmetric-keys to be allocated in this hierarchical key strategy, in a normal feeling.

We believe that there are exactly 2h cipher written text sessions, and the delegate of issue is eligible to a portion r of them. That is, r is the delegation rate, the ratio of the allocated cipher written text sessions to the complete sessions. Obviously, if r = 0, na should also be 0, which means no accessibility any of the classes; if r = 100%, na should be as low as 1, which indicates that the ownership of only the main key in the structure can allow the accessibility all the 2h sessions. Consequently, one may anticipate that na may first improve with r, and may reduce later [8]. We set r = 10%; 20%; …..; 90%, and select the section in a random way to design an irrelevant “delegation pattern” for different delegates. For each mixture of r and h, we arbitrarily produce 104 different combinations of sessions to be allocated, and the outcome key set size na is the common over unique delegations.

VI. EXPERIMENTAL SETUP

Our techniques allow the pressure aspect F (F = n in our schemes) to be a tunable parameter, at the price of O(n)-sized program parameter. Protection can be done in continuous time, while decryption can be done in O(jSj) team multiplications (or factor inclusion on elliptic curves) with 2 coupling functions, where S is the set of cipher text sessions decrypt able by the provided total key and jSj ¶ n [11]. As predicted, key removal needs O(jSj) team multiplications as well, which seems inevitable. However, as confirmed by the research outcomes, we do not need to set a very great n to have better pressure than the tree-based strategy. Observe that team multiplication is a very quick operate.

Depth of the Key	Time Efficiency
1	0.04985
2	0.05994
3	0.07012
4	0.08172
5	0.09860

Table1: Data processing with key structure with respect to time efficiency.

Again, we validate empirically that our research is real. We connected the essential KAC program in C with the Pairing-Based Cryptography (PBC) Library8 release 0.4.18 for the genuine elliptic-bend group and coupling capacities. Since the provided key can be as little as one G aspect, and the cipher text only contains two G and one GT components, we used (symmetric) combinations over Type-A (super singular) shapes as described in the PBC collection which provides the biggest performance among all kinds of shapes, even though Type-A shapes do not offer the quickest reflection for team components.

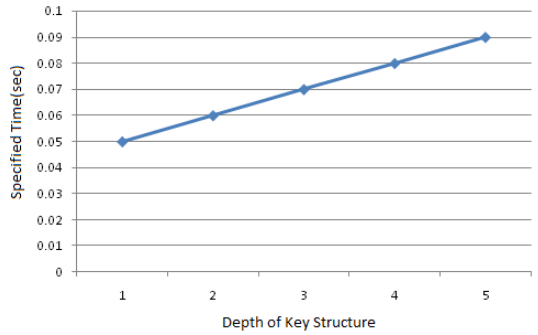


Figure 6: Experiments on program installation and top-level sector power allow. (a) Setup operation;

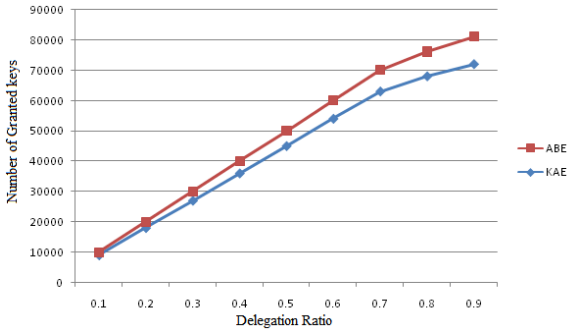


Figure 7: Variety of provided important factors (na) needed for different approaches in the situation of 65536 sessions of data.

The execution times of Installation, Key Gen, ensured are autonomous of the assignment rate r. In our tests, Key Gen requires 3:3 milliseconds and Protected requires 6:8 milliseconds. As anticipated, the working time complexities of Draw out and Decrypt enhance directly with the designation rate r (which chooses the measurement the doled out set S). Our minute results additionally agree to what can be seen from the equation in Draw out and Decrypt — two coupling capacities take insignificant time, the working length of time of Decrypt is around a double of Draw out. Watch that our tests took care of up to 65536 mixed bag of sessions (which is additionally the weight component), and ought to be sufficiently gigantic for fine-grained data examining as a rule [12]. In conclusion, we remark that for projects where the mixed bag of figure content sessions is colossal yet the non-private storage room is limited, one ought to set up our procedures utilizing the Type-D coupling included with the PBC, which just needs

170-bit to mean a component in G. For n = 216, the project parameter needs around 2:6 mb, which is as enormous as a lower quality MP3 data document or a higher-determination JPEG data record that a typical cellular telephone can shop more than various them. Be that as it may, we put away exorbitant secure storage room without the anxiety of taking care of a structure of assignment session.

VII. CONCLUSION

In this we show ABE for acknowledging adaptable, flexible, and fine-grained availability administration in thinking preparing. plan effortlessly has a progressive structure of framework clients by executing an assignment calculation to ABE not just encourages substance credits because of flexible list of capabilities blends, additionally finishes productive client crossing out on account of a few quality undertakings of components. The most effective method to secure clients' data solace is a primary inquiry of thinking storage room. With more measurable assets, cryptographic procedures are getting more adaptable and regularly incorporate a few imperative elements for one and only program. In this substance, we consider how to "pack" keys out in the open key cryptosystems which help designation of key essential elements for diverse figure content sessions in distributed storage. Whichever one among the force set of classes, the agent can simply get an aggregate key of constant measurement. Our methodology is more adaptable than various leveled key undertaking which can just protect spaces if every single key-holder talk about an indistinguishable arrangement of rights.

VIII. REFERENCES

[1] “Mohamed Nabeel, Elisa Bertino Fellow, “Privacy Preserving Delegated Access Control in Public Clouds”,” proceedings in A preliminary version of this paper appears in the Proceedings of the IEEE International Conference on Data Engineering(IRI ’12)[1] as an invited paper.

[2] “M. Nabeel and E. Bertino, “Privacy preserving delegated access control in the storage as a service model,” in *EEE International Conference on Information Reuse and Integration (IRI)*, 2012”.

[3] “N. Shang, M. Nabeel, F. Paci, and E. Bertino, “A privacy-preserving approach to policy-based content dissemination,” in *ICDE ’10: Proceedings of the 2010 IEEE 26th International Conference on Data Engineering*, 2010”.

[4] “M. Nabeel, E. Bertino, M. Kantarcioglu, and B. M. Thuraisingham, “Towards privacy preserving access control in the cloud,” in *Proceedings of the 7th International Conference on Collaborative Computing: Networking”, Applications and Worksharing*, ser. Collaborate Com ’11, 2011, pp. 172–180.

[5] “M.Nabeel, N.Shang, and E.Bertino, “Privacy preserving policy based content sharing in public clouds,” *IEEE Transactions on Knowledge and Data Engineering*, 2012”.

[6] “M. Nabeel and E. Bertino, “Towards attribute based group key management,” in *Proceedings of the 18th ACM conference on Computer and communications security*, Chicago, Illinois, USA, 2011”.

[7] “M.Nabeel and E.Bertino, “Attribute based group key management,” IEEE Transactions on Dependable and Secure Computing, 2012”.

[8] J.-M. Do, Y.-J. Song, and N. Park, “Attribute based proxy re-encryption for data confidentiality in cloud computing environments,” in *Proceedings of the 1st International Conference on Computers, Networks, Systems and Industrial Engineering*. Los Alamitos, CA, USA: IEEEComputerSociety,2011,pp.248–251.

[9] “Cheng-Kang Chu, Sherman S. M. Chow, Wen-Guey Tzeng, Jianying Zhou, and Robert H. Deng, “Key-Aggregate Cryptosystem for Scalable Data Sharing in Cloud Storage”, proceedings in This work was supported by the Singapore A*STAR project SecDC- 11217-2014”.

[10] “S. S. M. Chow, Y. J. He, L. C. K. Hui, and S.-M. Yiu, “SPICE - Simple Privacy-Preserving Identity-Management for Cloud Environment,” in Applied Cryptography and Network Security – ACNS 2012, ser. LNCS, vol. 7341. Springer, 2012, pp. 526–543”.

[11] “L. Hardesty, “Secure computers aren’t so secure,” MIT press, 2009, http://www.physorg.com/news176107396.html.
[12] “C. Wang, S. S. M. Chow, Q. Wang, K. Ren, and W. Lou, “Privacy- Preserving Public Auditing for Secure Cloud Storage,” IEEE Trans. Computers, vol. 62, no. 2, pp. 362–375, 2013”.
[13] “B. Wang, S. S. M. Chow, M. Li, and H. Li, “Storing Shared Data on the Cloud via Security-Mediator,” in International Conference on Distributed Computing Systems - ICDCS 2013. IEEE, 2013”.
[14] “S. S. M. Chow, C.-K. Chu, X. Huang, J. Zhou, and R. H. Deng, “Dynamic Secure Cloud Storage with Provenance,” in Cryptography and Security: From Theory to Applications - Essays Dedicated to Jean-Jacques Quisquater on the Occasion of His 65th Birthday, ser.LNCS, vol. 6805. Springer, 2012, pp. 442–464”.

Speed Optimised AES-GCM

N Rajitha [#1], R Sridevi [*2]

[#] Research Scholar, Department of Computer Science & Engineering
JNTUHCEH, Hyderabad, Telangana, India
[1]rajitha2k2@yahoo.co.in

[*] Professor, Department of Computer Science & Engineering
JNTUHCEH, Hyderabad, Telangana, India
[2]sridevirangu@jntuh.ac.in

***Abstract*—The Advanced Encryption Standard (AES) in Galois Counter Mode (GCM) has gained significant popularity as it is can be implemented in a parallelised and pipelined way particularly in the hardware. The multiplication operations in AES can be optimised using PCLMULQDQ as in the case of Intel Processors and also using Karatsuba Algorithm. This paper illustrates the existing works on optimisations of AES-GCM using Karatsuba Offman Algorithm.**

I. INTRODUCTION

Reconfigurable computing involves the use of reconfigurable devices, such as the field programmable gate arrays (FPGAs), for computing purposes [1]. Processing power is an important requirement not only for scientific computing applications but for interactive applications that include image processing, voice recognition, video streaming etc. Hence there is always demand on the computation units to implement these applications. One of the ways to support such huge processing requirement is to use reconfigurable computing systems.

A Reconfigurable computing system usually consists of one or more processors and a reconfigurable fabric [2]. The processor executes the sequential or noncritical code, while code that can be executed by processing units is mapped to the reconfigurable fabric. Furthermore, the parts of the computation that contain high degree of parallelism can be mapped to the reconfigurable unit.

AES-GCM is one of the algorithms that support green computing. GHASH and AES operations can be run in parallel and thus, it has been the choice for implementation.

Authenticated Encryption is a common requirement in many security applications as it guarantees both privacy and integrity. Several dedicated modes for authenticated encryption have been devised so far. The most popular such mode is the AES in Galois Counter Mode which is optimized for high performance[6].

Galois Counter Mode provides confidentiality, integrity and authenticity assurances on the data. Two functions in GCM are authenticated encryption (AE), authenticated decryption (AD) which includes GHASH. It can be efficiently pipelined and parallelized in hardware for implementation of these functions.

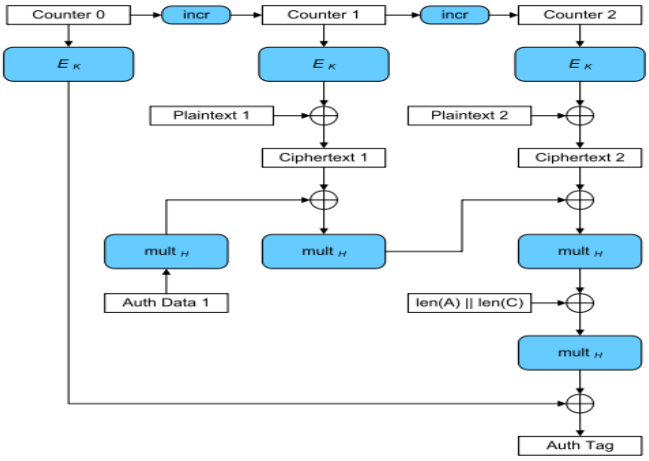


Fig. 1. AES-GCM Authenticated Encryption operation [7]

II. STEPS IN AES-GCM ALGORITHM

1. Let $H=E(K, 0^{128})$.
2. Define a block J_0, as follows:
   if len(IV) = 96, then let J_0 = IV || 0^{31} ||1.
   if len(IV) ≠ 96, then let s = 128 ⌈len(IV)/128⌉−len(IV), and let
   $J_0 = GHASH_H$ (IV||0s+64|| ⌈len(IV)⌉64).
3. $C=GCTR_k$ (inc32(J_0), P)
4. Let u=128 ⌈len(C)/128 ⌉−len(C)
   and let v=128 ⌈len(A)/128 ⌉−len(A).
5. Define a block, S, as follows:
   $S=GHASH_H$ (A || 0^u || C || 0^v || $[len(A)]_{64}$ || $[len(C)]_{64}$).
6. Let T = MSB_t ($GCTR_k$(J0, S)).

7. Return (C,T).

The steps involved in Authenticated Encryption function are as follows

1) Create an initial block 'H'.
2) Calculate a block J_0 based on the IV value and size.
3) Compute a block C such that $C=GCTR_k(inc_{32}(J_0), P)$.
4) Calculate values u, v to be used for padding, based on length of C, AAD
5) Calculate a block S using AAD, C, length of C, length of AAD and a padding with 0's with padding length equal to u+v
6) The most significant 't' bits of the result form Authentication Tag T

The following are the steps in Authenticated Decryption

1) Check for the length of IV, C, AAD are all supported else return FAIL
2) Create an initial block 'H'.
3) Calculate a block J_0 based on the IV value and size.
4) Compute a block C such that $C=GCTR_k(inc_{32}(J_0), C)$.
5) Based on the lengths of C and AAD calculate values u and v.
6) Calculate a block S using AAD, C, length of C, length of AAD and a padding with 0's with padding length equal to u+v
7) The most significant 't' bits of the result form Authentication Tag T
8) If T=T`, return(P). Else FAIL

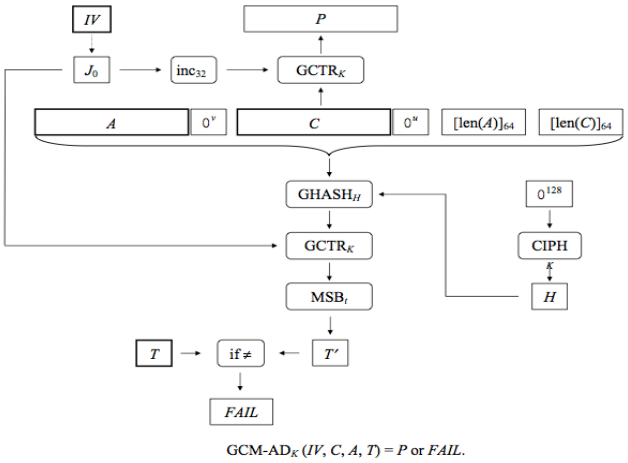
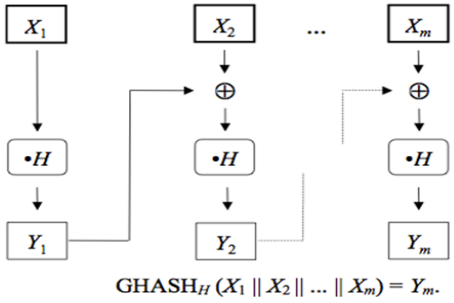


Fig. 2. AES-GCM Authenticated Decryption operation [7]

GHASH Function

The authentication mechanism within GCM is based on hash function, called GHASH (within a binary Galois field). The hash sub-key denoted as H, is generated by applying the block to the zero block. The GHASH is a keyed hash function but not, on its own, a cryptographic hash function. It is based on GF(2^{128}) multiplier with **irreducible polynomial**

$$\mathbf{F(x)=x^{128}+x^7+x^2+x+1}$$



GHASH is composed of chained GF(2^{128}) multipliers and bit wise XOR operations.

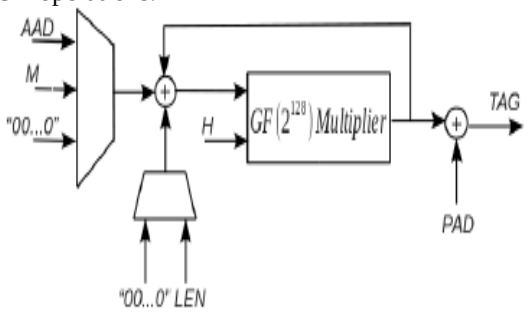


Fig. 3. GHASH Function[3]

The resulting 128-bit is expressed as
H= E(K,0128)
X0= GHASH(H,AAD)
X_i= GHASH(H,Mi⊕Xi−1)
LEN= length(AAD)64|| length(M)64
TAG= GHASH(H,Xn⊕LEN)
MAC= PAD⊕TAG

The GHASH architecture accepts 5 inputs:

- A 128-bit hash key H derived from a symmetric cryptographic key K.
- An M-bit message which can be divided into n 128-bit blocks M_1− M_n and the last message block M_n is padded with zeros to create a 128-bit word.
- A 128-bit Additional Authenticated Data (AAD) is authenticated but not encrypted.
- A 128-bit LEN value which expresses the word lengths of AAD and the message M.
- A128-bit cryptographic pad value(PAD) which ciphers the function output TAG to generate the message authentication code (MAC).

III. Parallel AES-GCM using Karatsuba Algorithm (KOA)

Integer multiplication algorithm takes $O(n^2)$ bit operations for multiplying two n-bit integers. A divide and conquer algorithm due to Karatsuba and Ofman reduces the

complexity to $O(n^{1.5})$. Karatsuba Ofman Algorithm (KOA) is used in [3] to reduce complexity of $GF(2^{128})$ multiplier.

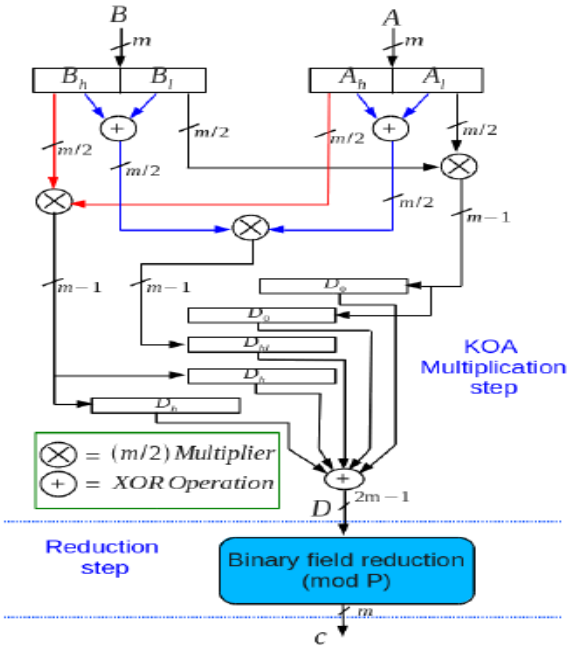


Fig. 4. KOA Multiplication [4]

$D_l = A_l B_l$
$D_{hl} = (A_h \oplus A_l)(B_h \oplus B_l)$
$D_h = A_h B_h$
$D = D_h X^m \oplus X^{m/2}(D_h \oplus D_{hl} \oplus D_l) \oplus D_l$

$X_i = (M_i \oplus X_{i-1}) \times H$
$= (M_i \times H) \oplus (X_{i-1} \times H)$
$= (M_i \times H) \oplus [(M_{i-1} \oplus X_{i-2}) \times H_2]$
$= M_i \times H) \oplus (M_{i-1} \times H_2) \oplus [(M_{i-2} \oplus X_{i-3}) \times H_3]$
$= (M_i \times H) \oplus (M_{i-1} \times H_2) \oplus (M_{i-2} \times H_3)$
$\oplus [(M_{i-3} \oplus X_{i-4}) \times H_4]$
$= ((M_i \times H) \oplus (M_{i-1} \times H_2) \oplus (M_{i-2} \times H_3)$
$\oplus (M_{i-3} \times H_4) .... \oplus [(M_N - 1 \oplus X_{N-2}) \times H_{N+1}]) \bmod P$

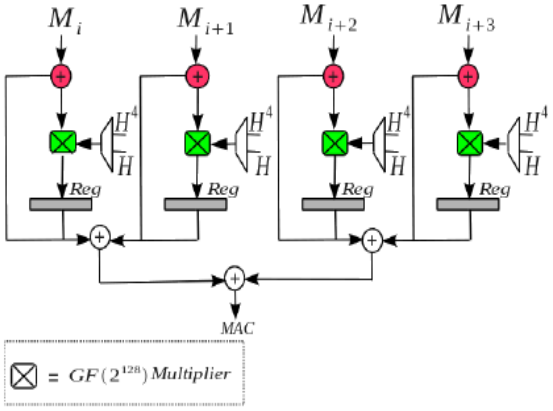


Fig. 5. Parallel GHASH Function[5]

$X_i = (X_{i-1} \oplus M_i)H$
$= (...(((M_1 H \oplus M_2)H \oplus M_3)H \oplus M_4)H...)H$
$= (((M_1 H_4 \oplus M_5)H_4 \oplus M_9)H_4 \oplus ...)H_4$
$\oplus (((M_2 H_4 \oplus M_6)H_4 \oplus M_{10})H_4 \oplus ...)H_3$
$\oplus (((M_3 H_4 \oplus M_7)H_4 \oplus M_{11})H_4 \oplus ...)H_2$
$\oplus (((M_4 H_4 \oplus M_8)H_4 \oplus M_{12})H_4 \oplus ...)H$

AES-GCM has been implemented in hardware to achieve high speeds with low cost and low latency in [10] GHASH fuction composed of chained $GF(2^{128})$ multipliers and bit wise XOR operations is implemented parallely. KOA was used to reduce the complexity of the $GF(2^{128})$ multiplier. Also, to reduce the data path of the KOA multiplier pipelining concept is used[8].

Constant key specialization on FPGA is used in [9] in which pre-computed keys are generated and synthesised into the architecture.

Recursive Karatsuba algorithm and generalization of polynomial multiplication for polynomials of degree 1 and degree N has been proposed in [11]. The following is a recursive Karatsuba algorithm.

Algorithm Recursive KA, C = KA(A, B)
INPUT: Polynomials A(x) and B(x)
OUTPUT: C(x) = A(x) B(x)
N =max(degree(A), degree(B))+1
if N ==1return A·B
Let $A(x)=A_u(x)\, x^{N/2}+A_l(x)$ and $B(x)=B_u(x)\, x^{N/2}+B_l(x)$
$D_0 = KA(A_l, B_l)$
$D_1 = KA(A_u, B_u)$
$D_{0,1} = KA(A_l+A_u, B_l+B_u)$
return $D_1 x^N + (D_{0,1} - D_0 - D_1)x^{N/2} + D$

IV. Uniqueness of Initialization Vector (IV)

The main design challenges to the implementation of cryptographic module implementing AES-GCM is the uniqueness requirement of initialization vector (IV) and freshness of the key [7]. It is assumed that any GCM key established among the users should be fresh with high probability and should be established with an standard key management protocol.

The secure operation of a cryptographic algorithm depends not only on the proper implementation of steps in the algorithm, but also on the adherence to the associated requirements. The authentication assurance in GCM crucially depends on the uniqueness of IVs. Also, it is possible that the hash subkey from the output i.e. cipher text can be extracted if the IVs ever get repeated for GCM and authenticated encryption functions with the same key.

IV can be constructed using two approaches Deterministic and Random Bit Generator (RBG) as per [7]. The IV in RBG based construction comprises of random field and free field. The total number of invocations of AE function shall not

exceed 2^{32} including the IV lengths. In RBG based construction, the above requirement along with the requirement that r(i)>=96 is enough to guarantee the uniqueness.

A major advantage of RBG based construction is that it can be designed to recover from loss of power without the intervention of the programmer. This may be achieved by including a non deterministic source of bits that are made available to the generator when the power is restored.

Extensive use of randomly generated bits is done in cryptographic applications. Nonces, one time pads (OTPs), Initialization Vectors (IVs), counters, keys are often used in security related applications. All these keying material is supposed to exhibit randomness. Besides, exhibiting randomness the keying material should also satisfy certain constraints and properties and to guarantee statistical distribution. Furthermore, they also need to pass through and fulfil tests of randomness to ensure non-repetition.

Till date a number of random number generators (RNG), have been developed which fall into two major categories true random number generators (TRNG), pseudo random number generators (PRNG).

PRNG based on irrational numbers is considered in [12]. Ring oscillator based random number generators on field programmable gate arrays (FPGAs) from Xilinx, Microsemi, Altera are evaluated in [13]. An extraction method for true random number generators which has high throughput Xilinx Spartan 6 FPGA is presented in [14].

V. Conclusion

Key synthesized AES-GCM is illustrated in [4] suitable for slow key changing applications like the Virtual Private Networks (VPNs). Similarly constant key and pre-computed are used in the previous works for implementing parallel and pipelined versions of AES-GCM on hardware to achieve high performance i.e., high throughput per slice.

References

[1] T J Todman et al, Reconfigurable Computing Architectures & Design methods, IEEE Proceeding Comput. Digit. Tech., Vol 152, No 2, March 2005

[2] R Tessier et al, Reconfigurable Computing Architectures, Proceedings of IEEE Vol103, No 3, Mar'15 (Invited Paper)

[3] H Fan et al, Obtaining more Karatsuba-like formulae over the binary field IET Inf. Secur., 2012, Vol. 6, Iss. 1, pp. 14–19

[4] Karim M. Abdellatif, R. Chotin-Avot, and H. Mehrez Improved Method for Parallel AES-GCM Cores Using FPGAs, IEEE 2013.

[5] Karim M. Abdellatif, International Conference on Reconfigurable Computing FPGA, Dec 2012.

[6] Heqen Chen, Authentication Encryption Modes of Block Ciphers their security & implementation properties, 2009

[7] SPD 800-38

[8] G. Zhou and H. Michalik, "Improving Throughput of AES-GCM with Pipelined Karatsuba Multipliers on FPGAs," Reconfigurable Computing:Architectures,ToolsandApplications,pp.193–203,2009.

[9] L. Henzen and W. Fichtner, "FPGA Parallel-Pipelined AES-GCM Core for 100G Ethernet Applications," Proceedings of the ESSCIRC, pp. 202–205, 2010.

[10] A. Satoh et al, High Speed Pipelined AES-GCM Core for 100G Ethernet Applications, ESSCIRC, 118-129, 2007

[11] Andr´e Weimerskirch and Christof Paar, Generalizations of the Karatsuba Algorithm for Efficient Implementations

[12] Luka Milinkovi´c,Marija Anti´c and Zoran ˇCiˇca, Pseudo-Random Number Generator Based on Irrational Numbers, Telsiks, IEEE Oct 2011

[13] MICHAEL RAITZA, MARKUS VOGT, CHRISTIAN HOCHBERGER, and THILO PIONTECK, Random Number Generatorson FPGAs, ACM Transactions on Reconfigurable Technology and Systems, Vol. 9,No. 2, Article 15,Publication date: December 2015

Vladimir Rozic1, Bohan Yang1, Wim Dehaene2, Ingrid Verbauwhede, Highly Efficient Entropy Extractionfor True Random Number Generators on FPGAs, DAC'15,June07 - 11,2015, ACM 978-1-4503-3520-1/15/06

Telugu numeral recognition using machine learning techniques

P.Harish, II M.Tech
Computer Science & Engineering,
VR Siddhartha Engineering College,
Kanuru, Vijayawada, Andhra Pradesh,
India.
harish.pattem2@gmail.com

Dr. S.Vasavi, Professor
Computer Science & Engineering
VR Siddhartha Engineering College
Kanuru, Vijayawada, Andhra Pradesh
India
vasavi.movva@gmail.com

***Abstract*—Recognition of characters and numerals has been always an important challenging task in image processing and pattern recognition. Recognizing characters and numerals from a given image is important in mainly applications such as number-plate recognition, postal cards, verification of identity cards etc...Adequate studies have been done on Chinese, Japanese, and Kannada etc...Even though some work have reported for Telugu numeral recognition, But there are no proper recognition systems for Telugu numerals in Telugu language. Existing method uses optical character recognition (OCR) which is the task of recognition of numerals that are present in a digital image, whose domain can be machine print or hand written. A method to recognize Telugu numerals from APSRTC number plates using template matching and machine leaning techniques like K-NN and SVM and construct number plates in English.**

Keywords: ***Machine learning techniques, Numeral-recognition, Optical-Character-Recognition, Template matching***

I. INTRODUCTION:

In Andhra Pradesh and Telangana states people use Telugu language for communication. Also union territories such as Yanam and Island such as Andaman & Nicobar use Telugu language. In Telugu language, Telugu numeral plays a vital role. Even now APSRTC buses are using Telugu numerals on the vehicle number plates. Telugu APSRTC bus numeral plate is shown in Figure 1.

Figure 1: APSRTC vehicle number plate using Telugu numerals.

Because of such usage, it is becoming problem for police authorities to recognize the vehicle number immediately in order to impose fine. As such, there is a need for identifying the Telugu numeral and converting into English numeral.

Figure 2 presents English numerals and its equivalent Telugu numerals.

0	1	2	3	4	5	6	7	8	9
౦	౧	౨	౩	౪	౫	౬	౭	౮	౯

Figure 2: Telugu numerals from 0-9

In the field of research, Image processing has been developing day by day, because of its applications in various fields like banking & security etc. Identification of numerals from a digital image is one part of the image processing. These can be applied in different fields like banking, postal, searching etc.

A. Optical Character Recognition

Optical character recognition that reads text from the paper and translating the image in the form human readable text to computer manipulate. EX: ASCII code. OCR contains optical scanner for reading text and elegant software for analyzing images. Most OCR use both software and hardware to recognize the characters. OCR can read the text in different fonts but hand written is still difficult.

B. K-Nearest Neighbour

For classification problems is more effective technique. Calculations can work through input patterns with the trained patterns. The K-NN classifier is based on the assumption that the

classification of an instance is most similar to classification other instances which are nearby to vector space.

C. Support Vector Machine

Support vector machine construct a hyper plane which separates positive and negative classes with a large margin. This margin classifier that solve quadratic programming program of minimum separation between two classes. The goal of SVM process for classification is maximize the distance between the separating lines and each data set.

In rest paper we discuss as follows: Section II about literature survey, Section III about proposed system, Section IV about Experimental results, finally paper end by Section V, conclusion and future work

II. LITERATURE SURVEY

In [1]—we are using 12 different rules to identify the Telugu numerals. Different methods like zoning method, cavity method and endpoint method are used for the identification of numerals. These rules used to identify the position of end points and checks whether cavities exist or not. Pre-processing is much needed for identifying the numerals. Then divide numerals into 2*2 zones and now those 12 rules can be applied to recognize the telugu numerals.

Advantages and disadvantages:

By using these simple we can easily identify printed Telugu numerals easily and efficiently. However, for hand written there rules are not enough. So it is needed to increase some more rules for them. Every time we do not identify the end points and cavities. By using some new rules, we can only identify the end points and cavity. So time complexity is reduce

In [2]—It uses to identify both hand written and printed numerals. Different methods are performing to identify printed telugu numerals.

A. Structural method

B. Skeleton method

C. water reservoir method

Different method are performed to identify the handwritten telugu numerals

A .Zoned based method

B. Moment invariant

In this preprocessing is much needed for identifying the numerals.

Advantage and disadvantages:

Comparative analysis can be provided between Telugu numerals and characters. Even though, many methods are proposed to recognize Telugu language. Still handwriting recognition is always a challenging task for Telugu language.

In [3]—for template matching we use convolution method. Template matching, in which combines input character with the trained character. In this work, convolution is used to identify the hand written numerals of Kannada by using pen, pencil, sketch and machine printed digits. Convolution methods differentiate 2000 numerals of different style, stroke, texture, thickness, thinness and fonts. Steps involved in this are:

1. Scan the image

2. Preprocess by using Binarization

3. Comparing input with template image

By comparing all these results sketch results is better compare to pen and pencil.

Database: Experiments were carried out with database containing 2000 handwritten Kannada digits there are 200 people of different age groups each of them written digits from 0 to 9.

Advantages and disadvantages:

It identifies the Kannada numerals wrote with pen, pencil, sketch and machine printed digits. Accuracy for pen, pencil, sketch and machine printed digits are 88.01, 82.36, 93.76,100 percent respectively

Disadvantage is to compare only these - 2000 samples in data base .If input given is not matching with the given 2000 samples then it doesn't shows accurately.

In [4]—uses composite method to identify handwritten Devanagari Numerals. In this stacking

approach is used to fuse the confidence of 4 different classifiers i.e.

1). Naïve Bayes
2) .Instance Based Learner
3).Random forest
4).Sequential minimal optimization (SMO).

It extracts both local and global features from the handwritten numerals. For features extraction we consider two sets of features.

1). Fourier Descriptor
2). Zonal features

Fourier Descriptor for global features and Zonal features for local features. It has been tested on large set of handwritten numeral database and gave accuracy of 99.685%.

Advantages and disadvantages:

Number of features used was very less and easy to compute. In this we take both global and local features through zonal and Fourier descriptor.

Disadvantage is that telling whether the given character is a numeral, vowel or consonant.

In [5]—it planned to design a quick and accurate Kannada numeral recognition. It uses different features. Few steps like pre-processing, segmentation, feature extraction and classification are involved in proposed numeral recognition. For pre-processing there are different functions are there to accurate results .Functions like

1) Thinning
2) Normalization
3) Noise removal

After the preprocessing, there have coded in 4 different ways of features there are

1) Density Feature
2) Left right profile
3) Number of crossover points
4) Detection of horizontal and vertical lines

Classification is done through nearest neighbor classifier.

Advantages and disadvantages:

Nudi kannada word processing software is used to generate printed kannada number. For particular fonts, this accurately identifies all the numerals. Density method is technique used to recognize the printed kannada numbers. This is for very good potential application. The size and fonts are independent. In this we have taken only 10 different fonts. Their work can be extended to other fonts too such as BRHKan01, BRHKan07 Bold Italic from Baraha and Nudifont' software.

In [6], for telugu script new method is construct using OCR system. Feature can extracted through

1). Singular Value Decomposition
2). Projection Profiles
3). Discrete Wavelet Transform

Classification can be done by using

1). K-Nearest Neighbor
2).Support Vector Machine

Dataset: For this we use Telugu Character database and CAMTE9Rdb are used. CAMTER db for Telugu numerals and Telugu character database for vowels consonants in telugu language.

Advantages and disadvantages:

Most capable results can be from DWT features with SVM classifier

Drawback is that, work can be extended by creating a good training Telugu character database consisting of all different font families and all characters existing for Telugu language.

III. PROPOSED METHOD

The proposed system of architecture is shown in Figure2

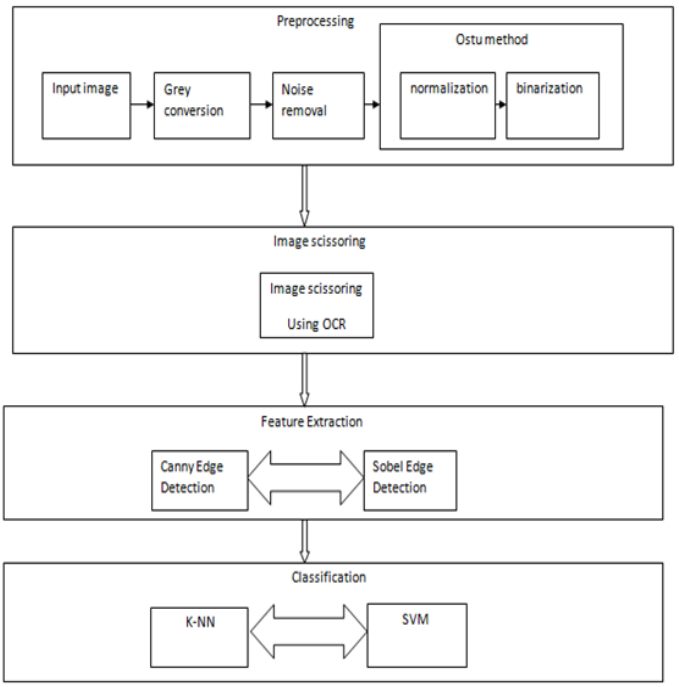


Figure2: Proposed system architecture

Figure 3 shows about the flow diagram of the architecture.

This Architecture consists of 2 phases:

1. Training Phase
2. Testing phase

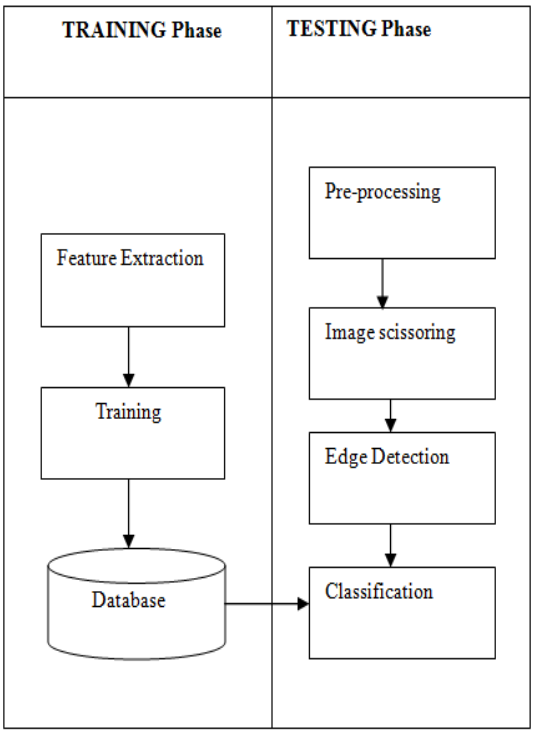


Figure 3: Architecture Flow diagram

Training phase:

Step 1: In this, we take database with 10 sets of Telugu numerals from 0-9
Step2: We train the 10 sets of database using 21 structural and texture features of the numerals using hue, saturation, values, and distances.
Step 3: Keep the 21features values in (dot) .csv file
Step 4: compare using SVM,K-NN classifier

Testing Phase*:*

Step 1: Take an input image
Step 2: Apply pre-processing step to input image
In Pre-processing step, we use

1. Noise removal
2. Min-Max Normalization
3. Binarization

2.1 Noise removal

We are using Median Filter for removal of noise from the given input image. By this we remove the noise and protect the edges of the numerals.

Formula for Median Filter:

Median[A(x)+B(X)]≠Median[A(x)]+ Median[B(x)]

2.2. Min-Max Normalization

We are using minimal and maximal normalization for normalization. By using this, light edge increases its thickness.

Formula for normalization:

$$I_n = (I - Min)\frac{NewMax - NewMin}{Max - Min} + NewMin$$

2.3.Binarization using Otsu method

We are using Otsu method for Binarization technique. Otsu converts into binary image.

Formula Ostu Method

$$\sigma_\omega^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t)$$

Weights $\omega_{0,1}$ are the probabilities of the two classes separated by threshold t and $\sigma_{0,1}^2$ are variances of these two classes.

Step 3: Applying image scissoring technique (OCR) for given Pre-processed image.

This can be done using following procedure

1. We should recognize the centre of matrix
2. By using distance formulae, Calculate radius by finding pixel with maximum distance from centre

$$Dist = \sqrt{((y2-y1)^2 + (x2-x1)^2}$$

3. Distinguish size of each track of imaginary
4. Recognize imaginary sectors.
5. Calculating number of 1's in each intersection of sector and track matrix.

Step 4: Apply edge Detection on scissoring image...i.e., Canny Edge Detection and Sobel Edge Detection

A. Canny Edge Detection:

The algorithm runs in 5 separate steps:

1. Smoothing: To remove noise of Blurring image
2. Finding gradients: The edges should be considered, where large magnitude of gradient of the image is

$$|G| = \sqrt{G_x^2 + G_y^2}$$

3. Non-maximum suppression: extreme points are noticed as edges.
4. Double thresholding
5. Edge tracking by hysteresis: The suppressing of all edges that are not connected to strong edge are determined as final stage.

B. Sobel Edge Detection:

The Sobel edge detector uses a pair of 3*3 masks, one estimating the gradient in the x-axis direction and other estimating the gradient in the y axis direction

Formula:

$$|G| = \sqrt{G_x^2 + G_y^2}$$

Step 5: Extract Texture and Structural Features of the image from Edge Detection image

Step 6: Apply classification on database with Trained Features with Test Features.
Classification can be done using

I. K-NN
II. SVM

ALGORITHM FOR PROPOSED METHOD:

INPUT : Telugu numeral
OUTPUT: English Numeral
Step 1: Take input image from APSRTC bus

Step 2: Apply Pre-processing step to that image

i) grey conversion
ii) Noise removal
iii) Normalization
iv) Binarization

Step 3: Apply image scissoring (OCR) to pre-processed image
Step 4: Apply edge detection techniques for scissor image.

i. Canny edge detection
ii. Sobel edge detection

Step 5: Apply structural and textual features to the edge detected image
Step 6: Apply classification techniques to the edge detected image and classify using

i. SVM
ii. K-NN

Code:

I. Pre-processing

1. Read image &Gray scale

```
Mat source = Highgui.imread("F:\\lptel\\New folder\\10.jpg",Highgui.CV_LOAD_IMAGE_GRAYSCALE);
```

2. Noise removal

```
Imgproc.filter2D (source, destination, -1, kernel);
```

3. Otsu method

```
sumB += (float) (t *histData[t]);

floatmB = sumB / wB;       // Mean Background

float mF = (sum - sumB) / wF;// Mean Foreground

// Calculate Between Class Variance

floatvarBetween = (float)wB * (float)wF * (mB - mF) * (mB - mF);
```

II. Segmentation (OCR)

```
if (blockX1 < 0)
    {       blockX1 = 0;
    }
else if (blockX1 >= w)
    {       blockX1 = w - 1;
    }
if (blockY1 < 0)
    {       blockY1 = 0;
    }
else if (blockY1 >= h)
    {       blockY1 = h - 1;
```

```
}
if ((blockX2 <= 0) || (blockX2 >= w))
{ blockX2 = w - 1;
}
if ((blockY2 <= 0) || (blockY2 >= h))
{ blockY2 = h - 1;
}
```

III. Feature extraction

1. Canny

```
fx = ( f(i-1,j-1) + f(i-1,j)/3 + f(i-1,j+1) ) - ( f(i+1,j-1) + f(i+1,j)/3 + f(i+1,j+1) );

fy = ( f(i-1,j-1) + f(i,j-1)/3 + f(i+1,j-1) ) - ( f(i-1,j+1) + f(i,j+1)/3 + f(i+1,j+1) );
```

2. Sobel

```
fx = ( f(i-1,j-1) + 2*f(i-1,j) + f(i-1,j+1) ) - ( f(i+1,j-1) + 2*f(i+1,j) + f(i+1,j+1) );

fy = ( f(i-1,j-1) + 2*f(i,j-1) + f(i+1,j-1) ) - ( f(i-1,j+1) + 2*f(i,j+1) + f(i+1,j+1) );
```

IV. Classification

1. *K-NN:*

```
Sample.Pixels [i-1] = double. Parse

Double (tokens[i]);
```

2. *SVM:*

```
clas.put(s[i],i);

System.out.println(clas.get(s[s.length-1]));

fileWriter.append(s[0].toString());
```

IV. EXPERIMENTAL RESULTS

Dataset: We are creating our own dataset. Each dataset contains handwritten and printed numerals. We divide each number into one class. All the 9 classes of numbers are kept in one database. We trained by SVM classifiers with structural& textual features. After classifying the input image is converted from Telugu numeral to English numerals as an output.

Database for the training sample are shown below the figure:4

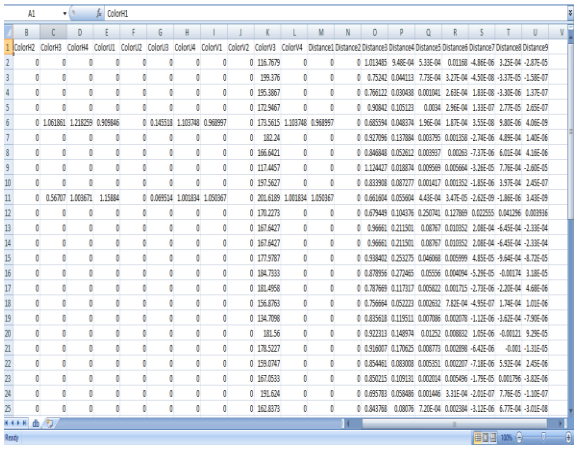


Figure: 4 Database for training sample

Database for the testing sample are shown below the figure: 5

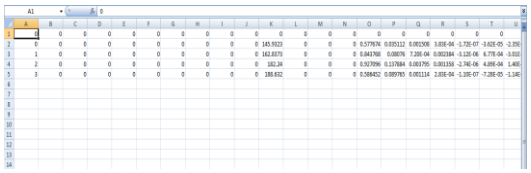


Figure: 5 Database for testing sample

We take input image of bus number. Input image will be shown in below Figure: 6

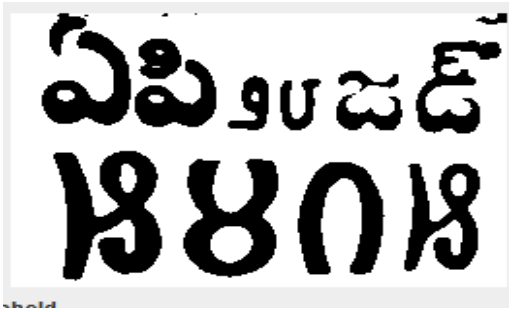


Figure 6: Bus number

When we apply preprocessing the result will be shown below the Figure: 7

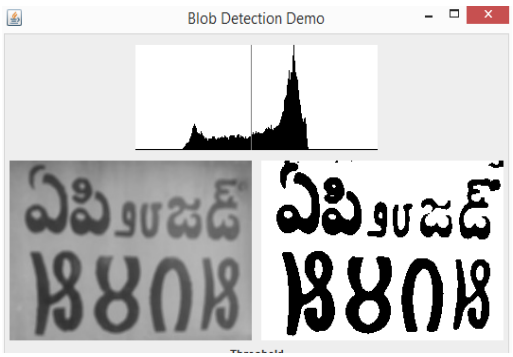


Figure 7: result after preprocessing

After the edge detection techniques the result be shown below Figure:8

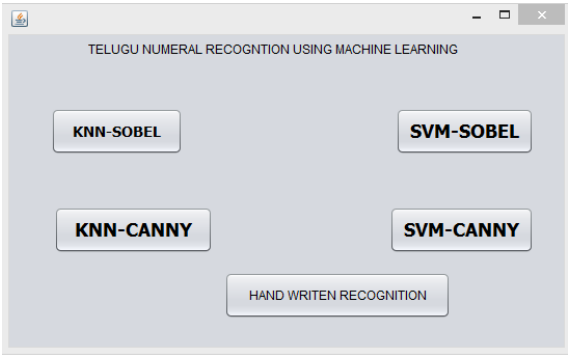


Figure: 8 Results for edge detection

Database values are saved in Csvfile that can be shown below the figure: 9

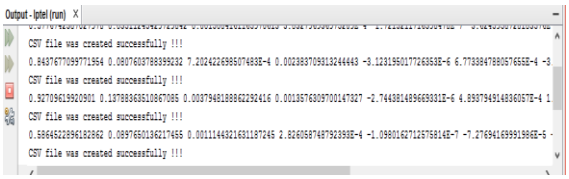


Figure 9: csv values creation

Result will be shown in below Figure:10

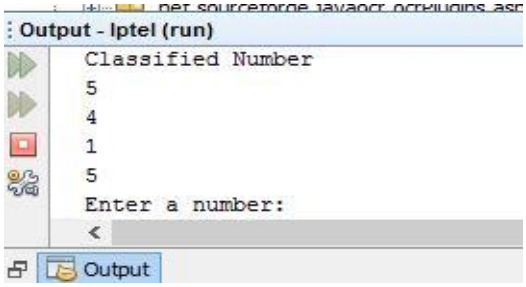


Figure 10: OUTPUT

Result for accuracy is shown in figure 11

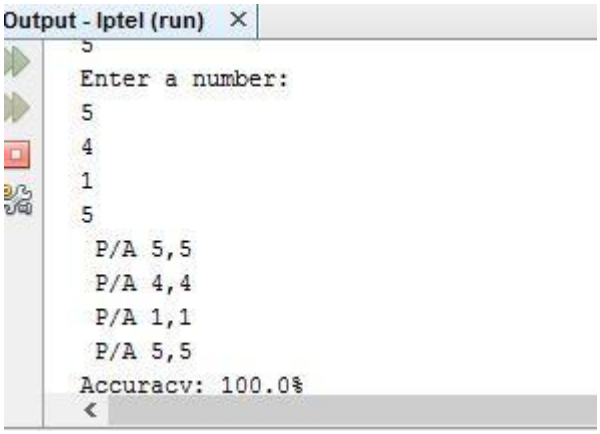


Figure 11 shows the accuracy of the output

Table 1 shows the Telugu number, converted English number, and recognization of numerals in K-NN&Canny, K-NN&sobel, SVM&Canny, SVM& Sobel

Telugu Bus number	*English bus number*	*K-NN Canny*	*K-NN Sobel*	*SVM& Canny*	*SVM& Sobel*
	5	Yes	yes	Yes	Yes
	4	Yes	yes	yes	yes
	1	Yes	yes	yes	yes
	5	Yes	yes	yes	yes

Table: 1

Conclusions and future work:

The method proposed here recognizes handwritten, as well as typewritten characters from the digital image. In this technique sobel texture, shape features were calculated on the image. After that, data undergoes classification process. The k-nearest neighbour classifier, SVM classifier were used to assign the unlabeled character object to a labelled class of character. For this purpose, a dataset has been used to train the classifier where the estimated values obtained from each cells are considered as the attributes for the objects. Various simulation results show that the proposed method can perform much accurately to recognize character. Besides, the proposed technique is less complex and

easy to implement while recognizing the characters from document image accurately as well.

For future work, large number of image datasets needs to be evaluated and tested with some more performance metrics. Future work aims to improve classifier to achieve still better recognition. Recognition is often followed by a post-processing stage. We hope and foresee that if post-processing is done, the accuracy will be even higher and then it could be directly implemented on mobile devices.

REFERENCES

[1].Ch. N. Manishaand, Y.K. Sundara Krishna, E. Sreenivasa Reddy,"Rule Based Recognition ofPrinted Telugu Numerals," Conference Paper Research Gate -Dec 2014.

[2] Ch. N. Manisha, E. Sreenivasa Reddy, Y.K. Sundara Krishna "A Study On Recognition Methods Of Telugu Numerals And Characters," International Journal of Emerging Technology in Computer Science & ElectronicsNov 2014.

[3] ShivanandKilledar, SatishDeshapande,"Kannada Handwritten Numerals Recognition And Translation Using Template Matching," International Journal on Recent Technologies in Mechanical and Electrical Engineering June 2015.

[4] Prabhanjan S andR Dinesh,"Handwritten Devanagari Numeral Recognition by Fusion of Classifiers,"International Journal of Signal Processing, Image Processing and Pattern Recognition- 2015.

[5]ShantalaShettar,BasavaprasadB,Smt. Bhagya H. K, "Recognition Of Printed Kannada Numerals By Nearest Neighbor Method," International Journal of Innovative Technology and Research.- April 2015.

[6] J. Jyothi, K. Manjusha, M. Anand Kumar and K. P. Soman," Innovative Feature Sets For Machine Learning Based Telugu Character Recognition,"Indian Journal of Science and Technology -2015.

AUTHOR(S) PROFILE

P. Harish has received his B.Tech Degree in Computer Science and Engineering, Vijayawada. His research includes Digital Image Processing and Pattern Recognition currently pursuing M.Tech in Computer Science and Engineering from VR Siddhartha College (Autonomous), Vijayawada, India.

Dr. S. Vasavi is working as a Professor in Computer Science & Engineering Department, VR Siddhartha Engineering College Vijayawada, India. With 19 years of experience. She currently holds funded Research projects from University Grants Commission (UGC) and Advanced Data Research. Institute (ADRIN, ISRO). Her research areas are Data mining and Image Classification. She is a life member of Computer Society of India and Published 35 papers in various International conferences and journals.

Exploring Spectral Features for Emotion Recognition Using GMM

J. Naga Padmaja, R. Rajeswar Rao

Computer Science and Engineering, JNTU Hyderabad
Khammam, INDIA
srija26@gmail.com

Computer Science and Engineering, JNTU Vizianagaram
Vizizanagaram, INDIA
raob4u@yahoo.com

***Abstract*—Automatic Text-Independent Emotion Recognition from Speech is a system which identifies the particular emotion automatically without basing on any particular text or a particular speaker. Emotion Recognition from speech is particularly useful for applications in the field of human machine interaction to make better human machine interface.It is used as lie detector and voice tag in different database access systems,in telephone shop, ATM machine as a password for accessing the particular account.It was given that Emotion Recognition from speech is classified into two types which are automatic text independent emotion recognition which has the same text in training database and testing database and Automatic Text Dependent emotion recognition has the different text in the training and testing database.**

An important step in emotion recognition from speech is to select a significant features which carries large emotional information about the speech signal, it was given that speech signal has different types of features among them are prosody, spectral and acoustic features. In this the Spectral features are used such as MFCC, Spectral Centroid , Spectral Skewness and Spectral Pitch Chroma. These features have been modeled by Gaussian mixture model and optimalnumber of Gaussians are identified. The database which was used in this system is Telugu database(IITKGP-Simulated Emotion Speech corpus) and four emotions are considered which are Anger, Fear, Neutral and Happy. Finally the experiments were conducted on different combinations of spectral features and it is established that the combination of MFCC and Skewness gives the better accuracy comparing all the combination of spectral feature.

Keywords: MFCC, Spectral Centroid , Spectral Skewness and Spectral Pitch Chroma, GMM

INTRODUCTION

An emotion is a human state of a mental behavior which expresses the feeling by physical moments or by words. As the physical moments are the facial expression and the body language, words are the way they speak and the way they pronounce the words. People express emotions as part of everyday communication. Emotions can be judged by a combination of cues such as facial expressions, prosodies, gestures, and actions.

Emotion recognition based on a speech signal is one of intensively studied research topics in the domains of human-computer interaction and affective computing. This research aims to evaluate the potential for emotion recognition technology to improve the quality of human computer interaction. The specific objectives of this research To establish the extent to which people will naturally express emotions when they know they are interacting with an emotion-detecting computer, To identify the conditions under which the application of emotion detection can lead to improvements in subjective and/or objective measures of system usability, To provide Human Factors guidelines on the deployment of emotion recognition technology which can help the developers of such technology to meet the needs of real users.

Emotion recognition from the speaker's speech is very difficult because of the following reasons:

In differentiating between various emotions which particular speech features are more useful is not clear. Because of the existence of the different sentences, speakers, speaking styles, speaking rates accosting variability was introduced, because of which speech features get directly affected. The same utterance may show different emotions. Each emotion may correspond to the different portions of the spoken utterance. Therefore it is very difficult to differentiate these portions of utterance. Another problem is that emotion expression is depending on the speaker and his or her culture and environment. As the culture and environment gets change the speaking style also gets change, which is another challenge in front of the speech emotion recognition system. There may be two or more types of emotions, long term emotion and transient one, so it is not clear which type of emotion the recognizer will detect.

It is a complex task that is furthermore complicated by the fact that there is no unambiguous answer to what the "correct" emotion is for a given speech sample. The vocal emotions explored may have been induced or acted or they may be have been elicited from more "real", lifelike contexts. Spontaneous speech from actual telephone services could be counted as such a material. The line of emotion research can roughly be viewed as going from the analysis of acted speech to more "real". The motivation of the latter is often to try to

enhance the performance of human-machine interaction systems, such as voice controlled telephone services.

The important issues in speech emotion recognition system are the signal processing unit in which appropriate features are extracted from available speech signal and another is a classifier which recognizes emotions from the speech signal. The average accuracy of the most of the classifiers for speaker independent system is less than that for the speaker dependent. The complexity of the task of automatic emotion recognition from speech increases with the naturalness of the assets—the recognition of natural emotions is much more challenging than that of acted ones.

Speech is the most natural form of human communication. Speech is one of the most information-laid signals; speech sounds have a rich and multi-layered temporal-spectral variation that convey words, intention, expression, intonation, accent, speaker identity, gender, age, style of speaking, state of health of the speaker and emotion. Speech sounds are produced by air pressure vibrations generated by pushing inhaled air from the lungs through the vibrating vocal cords and vocal tract and out from the lips and nose airways.

The air is modulated and shaped by the vibrations of the glottal cords, the resonance of the vocal tract and nasal cavities, the position of the tongue and the openings and closings of the mouth. Speech signal contain information like intended message, speaker identity and emotional state of speaker. An important issue in speech emotion recognition is to determine a set of important emotions to be classified by an automatic emotion recognizer.

1. SPEECH EMOTION RECOGNITION SYSTEM

Emotion Recognition from Speech mainly has three working steps which are Feature Extraction, Training and Testing.

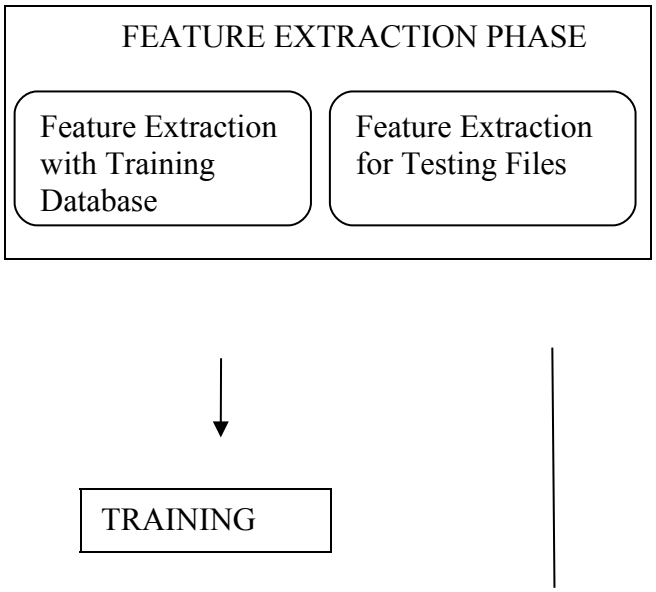


Fig 3.1: Pattern Recognition Task

Emotional speech recognition having in mind three goals. The first goal is to provide an up-to date record of the available emotional speech data collections. The number of emotional states, the language, the number of speakers, and the kind of speech are briefly addressed. The second goal is to present the most frequent acoustic features used for emotional speech recognition and to assess how the emotion affects them. Typical features are the pitch, the formants, the vocal tract cross-section areas, the Mel-frequency cepstral coefficients, the Teager energy operator-based features, the intensity of the speech signal, and the speech rate. The third goal is to review appropriate techniques in order to classify speech into emotional states. We examine separately classification techniques that exploit timing information from which that ignore it. Classification techniques based on hidden Markov models, artificial neural networks, linear discriminant analysis, k-nearest neighbours, support vector machines are reviewed.

In Shashidhar G. Koolagudi et all [1], the impulse like signal caused due to vocal folds' closure, within a pitch period is known as an 'epoch'. Though entire glottal pulse is responsible for the excitation of vocal tract, significant excitation takes place at epoch locations. So epoch parameters play an important role in any of the speech tasks.

In Iliou et al [2] presents an emotion recognition framework based on sound processing could significantly improve human computer interaction. One hundred thirty three (133) speech features obtained from sound processing of acting speech were tested in order to create a feature set sufficient to discriminate between seven emotions.

In Iker Luengo et al [4] very first experience, only pitch and energy related features were used. Intonation and energy curves are extracted from the recordings, both in linear and logarithmical scale. First and second derivative curves are calculated, as the pitch and energy change rate may provide new useful information for the recognition.

In K. Sreenivasa Rao et al [5], emotion recognition (ER) systems are developed using local and global prosodic features, extracted from sentence, word and syllable levels.

Word and syllable boundaries are identified using vowel onset points (VOPs) as the anchor points (Vuppala et al. 2012). In this work, VOP detection is carried out using the combination of evidence from excitation source, spectral peaks, and modulation spectrum. SVM (Support Vector Machine), Neural Networks, Decision trees are employed and for the vectors of short-term features HMM (Hidden Markov Model) is used for its dynamic performance.

In the above approaches they have used the standard basic feature extraction techniques, We proposed the combination of MFCC and spectral features which carries large information about the speech signal.

2. FEATURE EXTRACTION

One of the most important parts of emotion recognition from speech systems is the feature extraction process, selecting the right features is crucial for successful classification. Speech signal composed of large number of features which indicates emotion contents of it, changes in these features indicate changes in the emotions. Therefore proper choice of feature vectors is one of the most important task.

The spectral features play an important role in Speech emotion recognition. In stressed speech, the vocal tract spectrum is modulated resulting changes in overall spectrum. The spectral shape features are the features computed from short time fourier transform of the signal. A frame-by- frame analysis is performed using window size 20ms and shift 10ms.

2.1 MFCC:

Mel Scale Frequency Cepstral Coefficients represents the spectrum with few efficient which are called as frequency components. The cepstrum is the Fourier transform of the logarithm of the spectrum. Mel Frequency Cepstral Coefficients (MFCC) is commonly used as feature extraction technique in Emotion recognition system such as the system which can be automatically recognize which is the task of recognition emotion from their voice .MFCC are also increasingly finding uses in musicinformation such as gender classification, audio similarity measure etc.

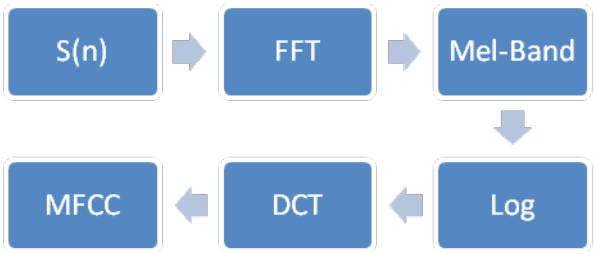


Fig 2.1: Steps for extraction of MFCC features

In this, continuous speech signal blocked into small frames of n samples, with next frames separated by m samples (m<n) with this the adjacent frames are overlapped by n-m samples.Windowing is done for minimizing the disruptions at the starting and at the end of the frame. Many windows represented with rectangular box.FFT is used for doing conversion from the spatial domain to frequency domain. Each frame have n samples are converted into frequency domain.The above calculated spectrum are mapped on Mel-scale to know the approximation about the existing energy at each spot with the help of triangular overlapping window also known as triangular filter bank.

$$\text{Mel}(f) = 2595 * \log 10\,(1 + f/700)$$

The process of carrying out DCT is done in order to convert the log Mel spectrum bank into the spectral domain.

2.2 SPECTRAL CENTROID:

Spectral Centroid is a good predictor of the brightness of sound. It is used in digital audio signal processing, it is a measure used in digital signal processing to characterize a spectrum. It is calculated as weighted mean of the frequencies present in the signal determined using FFT with their magnitudes and weights.

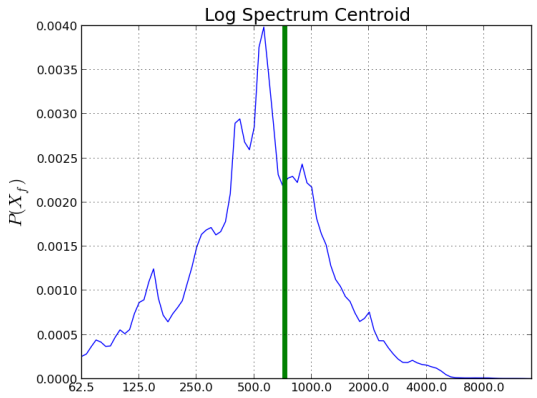


Fig 2.2: Spectral centroid on a Spectrum

The main advantage of Spectral centroid is it can detect the approximate location of formants.

Spectral Centroid=$\frac{\sum k\, F[k]}{\sum F[k]}$ {k=1 to N}

In practice, Centroid finds this frequency for a given frame, and then finds the nearest bin for that frequency. The centroid is usually a lot higher than one might intuitively expect, because there is so much more energy above the fundamental which contributes to the average.

2.3 SPECTRAL SKEWNESS:

The Skewness is a measure for how much the shape of the spectrum below the Centre of gravity is different from the shape above the mean frequency.

For a white noise, the Skewness is zero. Thespectral center of gravity is a measure for how high the frequencies in a spectrum are on average.

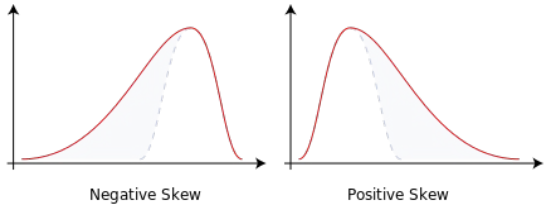


Fig 2.3: Spectral Skewness on a Spectrum

The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be Left- Skewed.

A Negative Skewness indicates more energy on the lower part of the spectrum. The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be Right-Skewness.
A Positive Skewness indicates more energy on the high frequency of the spectrum.
The main advantage of Skewness is, it shows the energy of a spectrum based on positive and negative Skewness.

For a sample of *n* values, The Skewness is

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}},$$

Where, $\overline{x}$is the Sample mean, s is the sample standard deviation, and the numerator m_3 is the sample third central moment.

2.4 PITCH CHROMA:

Chroma describes the angle of pitch rotation as traverses the helix.

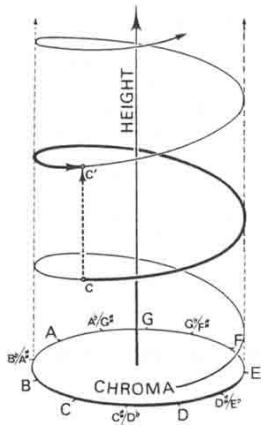


Fig 2.4: Pitch Chroma

Two octave-related pitches will share the same angle in the Chroma circle, a relation that is not captured by a linear pitch scale. It describes the angle of pitch rotation, Chroma features are robust to noise and loudness.

A pitch class is a set of all pitches that are a whole number of octaves apart, e.g., the pitch class C consists of the Cs in all octaves. "The pitch class C stands for all possible Cs, in whatever octave position." Thus, using scientific pitch notation, the pitch class "C" is the set $\{C_n: n \text{ is an integer}\} = \{..., C_{-2}, C_{-1}, C_0, C_1, C_2, C_3 ...\}$;

Although there is no formal upper or lower limit to this sequence, only a limited number of these pitches are audible to the human ear. Pitch class is important because human pitch-perception is periodic: pitches belonging to the same pitch class are perceived as having a similar "quality" or "color", a property called octave equivalence.

3. CLASSIFICATION

Gaussian Model is a probabilistic model for density clustering and estimation. GMM's are very efficient in modelling multi-model distributions. GMM is based on the assumptions that all vectors are independent. GMMs are used as classification tools to develop emotion recognition models. Mixture models ae a type of density model which comprise a number of component functions, usually cause. These component functions are combined to provide a multi model density. Determining the optimum number of Gaussian components is an important but theoretically difficult problem. We have explored different number of Gaussian components such as 2, 4, 8, 16, 32, 64, 128 and 256.

The Gaussian probability density function in one dimension is a bell shaped curve defined by two parameters, mean μ and variance σ^2. In the D-dimensional space it is defined in a matrix form as

$$N(x; \mu; \sum) = \frac{1}{(2\pi)^{D/2} |\sum|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu)^T \sum^{-1} (x - \mu)\right]$$

4. SPEECH DATABASE

The speech corpus, IITKGP-SESC, used in this study, was recorded using 6 (3 male and 3 female) professional artists from All India Radio (AIR) Vijayawada, India. The artists were sufficiently experienced in expressing the desired emotions from the neutral sentences. All the artists are in the age group of 25–40 years, and had the professional experience of 8–12 years. For analyzing the emotions we had considered 15 semantically neutral Telugu sentences. Each of the artists had to speak the 10 sentences in 4 given emotions in one session. The number of sessions considered for preparing the database was 10. The total number of utterances in the database was 2400 (10*sentences* × 4*emotions* × 6*artists*× 10 *sessions*). Each emotion had 640 utterances. The number of words and syllables in the sentences were varying from 3–6 and 11–18 respectively. The total duration of the database was around 36 minutes. The four emotions considered for collecting the proposed speech corpus were: Anger, Fear, Happiness, Neutral.The training duration 30 seconds and the testing duration of 3 seconds which was considered to test the system. The speech samples were recorded using SHURE dynamic cardioid microphone C660N. The distance between the microphoneand the speaker was maintained approximately around 3–4 inches. The speech signal was sampled at 16 kHz, and each sample is represented as 16 bit number. The sessions were recorded on alternate days to capture the inevitable variability in the human vocal tract system. In each session, all the artists have given the recordings of 10 sentences in 4 emotions. The recording was carried out in such a way that each artist had to speak all the sentences at a stretch in a particular emotion. This provides the coherence among the sentences for each emotion category. The entire speech database was recorded using single microphone and at the same location. The recording was done in a quiet room, without any obstacles in the recording path.

5. EXPERIMENTAL RESULTS

Comparing with all features which are considered in this system, the combination of MFCC and Spectral Skewness gives the better accuracy among all the combinations with the MFCC features. As the Spectral Skewness depends on the shape of the spectrum which indicates the energy on the higher or lower parts of the spectrum, it is a measure of how much the center of gravity is different from the shape above the mean frequency, the center of gravity is a measure of how high frequencies in a spectrum are on average.

Details:

M – Male

F - Female

30 sec Training and Testing Database

different classifications like SVM, HMM will be studied to evaluate the Emotion Recognition performance.

The below table explains multiple combinations with MFCC and the mixtures are CENTROID, SKEWNESS, PITCH CHROMA.

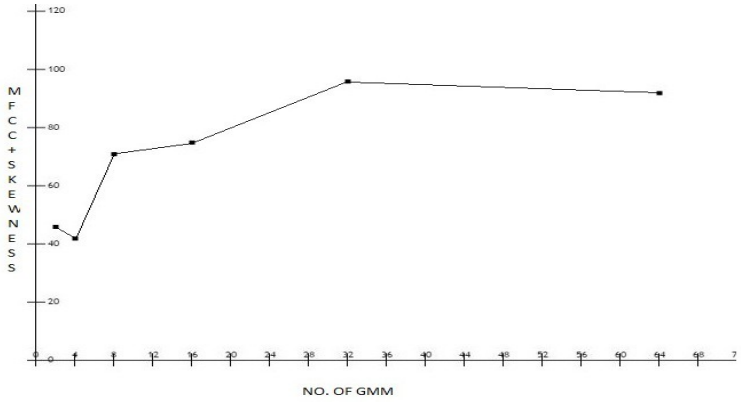
The figure (a) explains multiple mixtures and their bar graph representations. Here NO. Of GMM are taken on x-axis and mixture percentages are taken on y-axis.

The bar graphs with mixtures as M, M+C, M+S, M+PC, M+C+S, M+S+PC, M+C+PC, M+C+PC+S are taken and represented based on the NO.OF GMMs.

No of GMM	MFCC	MFCC+ Centroid	MFCC+ Skewness	MFCC+ PitchChroma	MFCC+ Centroid+ Skewness	MFCC+ Skewness+ PitchChroma	MFCC+ Centroid+ PitchChroma	MFCC+ Centroid+ Skewness+ PitchChroma
2	38%	33%	46%	33%	38%	33%	29%	25%
4	42%	46%	42%	50%	46%	54%	63%	46%
8	79%	67%	71%	63%	50%	58%	71%	54%
16	74%	88%	75%	83%	67%	83%	71%	67%
32	88%	83%	96%	83%	88%	88%	88%	83%
64	72%	46%	92%	88%	88%	92%	71%	92%
N	N=32	N=16	N=32	N=64	N=32	N=64	N=32	N=64

Figure (a)

M-MFCC , C-CENTROID, S-SKEWNESS, PC-PITCH CHROMA



Figure(b)

No of GMM and MFCC with SKEWNESS are taken on X and Y-axis respectively. Here at point 32, it occurred as maximum value with 96% which is shown in figure (b).

6. REFERENCES

[1] Koolagudi. S.G, Reddy. R, Rao. K.S, "Emotion recognition from speech signal using epoch parameters", IEEE Signal Processing and Communications (SPCOM), 2010 International Conference, ISBN: 978-1-4244-7137-9, Pages: 1-5, July 2010.

[2] Iliou, Anagnostopoulos., "Statistical Evaluation of Speech Features for Emotion Recognition", IEEEDigital Telecommunications, 2009. ICDT '09. Fourth International Conference, ISBN:978-0-7695-3695-8, pages: 121 – 126, July 2009.

[3] Dmitri Bitouka, RaginiVermaa, AniNenkovab, "Class-level spectral features for emotion recognition", Speech Communication, Volume 52, Issues 7–8, July–August 2010, Pages 613–625.

[4] Iker Luengo, Eva Navas, InmaculadaHernáez, Jon Sánchez, "Automatic Emotion Recognition using Prosodic Parameters"

[5] Shashidhar G. Koolagudi, Rao SreenivasaKrothapalli, Ramu Reddy Vempada"Emotion recognition from speech using source, system and prosodic features". Springer Int J Speech Technol (2013) 16:143–160, DOI 10.1007/s10772-012-9172-2

[6]S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1999.

ISBN 0-12-686140-4.

[7] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. John Wiley & Sons, Inc., 2nd edition, 2001.

[8] J.-K. Kämäräinen, V. Kyrki, M. Hamouz, J. Kittler, and H. Kälviäinen. Invariant Gabor features for face evidence extraction. In Proceedings of the IAPR Workshop on Machine Vision Applications, pages 228–231, Nara, Japan, 2002.

[9] J.-K. Kämäräinen. Face evidence extraction using Gabor features. Website. [Re- trieved 30.10.2003] From: http://www.it.lut.fi/project/facedetect/.

[10] S. Theodoridis and K. Koutroumbas. Pattern Recognition. Academic Press, 1999.

ISBN 0-12-686140-4.

[11] B.S. Everitt and D.J. Hand. Finite Mixture Distributions. Monographs on Applied Probability and Statistics. Chapman and Hall, 1981.

[12] J. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, 1997.

[13] M.A.T. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):381–396, Mar 2002.

[14]. D.S.Shete "Zero crossing rate and Energy of the Speech Signal of Devanagari Script", IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 4, Issue 1, Ver. I, PP 01-05, 2014

[15]. DimitriosVerveridis, Constantine Kotropoulos "Emotional speech recognition: Resources, features, and methods",

Perceptive Reckon System with RFID-tag with IEEE 802.15.4 technology

Azeem Mohammed Abdul[#1], Syed Umar[#2]

[# 1]*M.tech Student, Department of Electronics and Communication Engineering, KL University, Vaddeswaram.*

[#2]*Professor, Department Of Computer Science Engineering, GIST, Jaggayyapet*

[#1] *mohammedazeem123@gmail.com* [# 2] umar332@gmail.com

Abstract: Design based on microcontrollers, has won the status of the liveliest areas of electronics. This is a highly specialized area that is integrated on a single silicon chip, the power of thousands of transistors. At this time in the mall to buy some items trolley. As to know that the audience is a shopping centre in the metro cities. In particular, people buy a lot of products in the shopping and put them in the basket. In the case of the tax is collected scan barcodes. The process takes a long time. To prevent this, we have a system that we call intelligent wireless billing cart. This system makes use of RFID tags to replace the scanner. RFID tags have the product. Each time the product basket and the customer that you RFID reader and the prices of the goods and the cost scans on the LCD screen. As this process continues. We will send ZigBee is a vehicle used to transfer data to the host computer. The receiver has more information about your computer ZigBee transmitter. ARM7 core circuit (LPC2148), NXP Semiconductors (Philips) of the IC. The numbers in the LCD 16x2 used. It is used to display the product name, product costs, etc.

Keywords: RFID-tag, Zigbee, Billing System, Barcode scanner.

1. INTRODUCTION

Bar-codes have for years and are used by department stores and supermarkets to track buy control of goods from customers and inventory. However, the system not mean that the best way for business. Customers tired of waiting long, slow his hand in the transaction, especially on holidays. The cost reduction and mass production technologies of semiconductors for moving objects, looking for new markets where they can use a semiconductor chip. The result of the use of RFID also known as smart labels. RFID stands for Radio Frequency Identification. Smart Cart is equipped with Radio Frequency Identification (RFID) to identify the catalog server. Moreover, it also has an LCD screen that users export prices, discounts, deals and let the total weight. Once that thing or fell from the basket, the RFID tag identifies the account of products and updates. When customers shopping is finished, press the "business purpose" and the information is sent to the storage server and the customer only pays you the amount and suggested leave the cart easy to use and requires no training unique. Built-in automatic system for collecting and this system makes shopping a breeze and other positive effects, such as the release of personnel entering repetitive side reduce theft and increase efficiency in the matter.

2. ONGOING BILLING SYSTEM IN MARTS

2.1 The traditional methods of collecting

At present, access to the bar code as a shopping center. As QR barcode reader to read all barcodes product which scanner or barcode reader is a cover to electronic devices of a light source, a lens and a sensor light translation of optical pulses to the device. Moreover, the image sensor contains almost all readers' bar code decoder circuit test data and sends the contents of the barcode scanner color prints. If you choose to purchase a product from a basket and bring it to pay at the counter. The cashier scans show the barcodes of the products and the readers to give us an account. But a slow process when scanning a lot of products, so the billing process slow. This leads to the end of long queues.

2.*2 Barcode scanner vs. RFID*

For comparison, the RFID technology has been proven more durable than the barcode technology. You can read RFID tags from a distance. RFID reader, the information on the label and a distance of about 300 feet, but the technology QR code cannot be read by more than 15 feet away. RFID and bar code technology for increased speed. RFID tags can be viewed as more of a bar code. Barcode scanners are relatively conservative, because the line of sight required. On average, the second bar code readers of two pins, yet the RFID reader 40. RFID tags are well protected or can be on the machine, so that is not exposed to extreme wear. The interpretation of the barcode requires a direct view of the printing of bar codes, bar code printed on the external device and the most affected. Moreover, limited reuse of code bars. As for the barcode surgical skills, you cannot increase the information available about it. On the other hand, it is possible RFID.

3. INTELLIGENT BILLING SYSTEM

3.1 Block Diagram

Environmental technical wagon collection system, each has a removable radio frequency ID tag with a unique electronic product door. BX_CODE Met Electronic Product Code Information such as name, price, etc. of the product. When customers pick up the products from the shopping center technology, Radio Frequency ID tags and numbers called Electronic Product Code Radio Frequency Identification scan player. RF ID reader Electronic Product Code Arm 7 micro-control with 7 database system electronic product where a variety of products. After the price for their products is displayed on the LCD display technology carts for the collection, where users can view product information. Mobile 7 micro-controller for the results obtained from the database and ZigBee transmitter, wherein the data wireless transmission to a computer to load data. Computer data through Max 323. Max 323 port interface ZigBee receiver is connected based on the ZigBee computer accepts

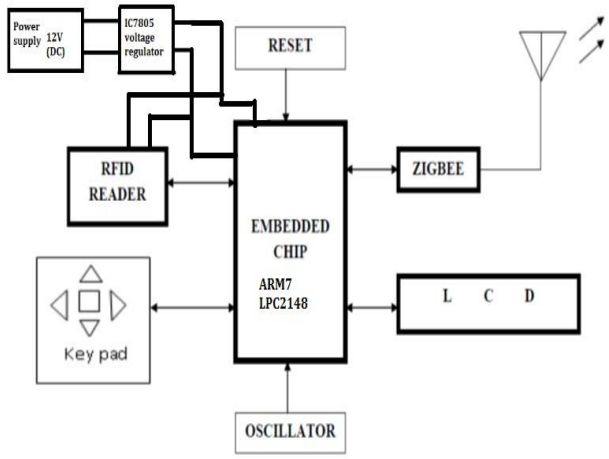


Fig.1: Block Diagram of the Trolley Section

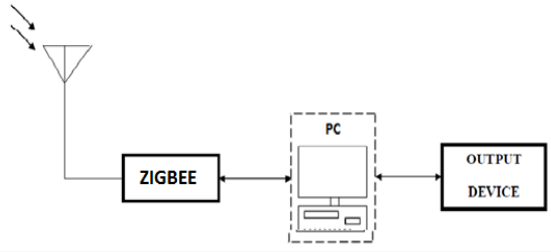
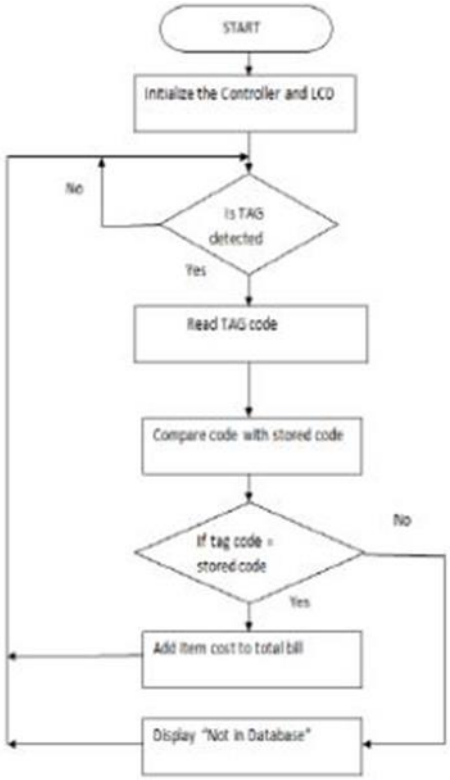


Fig.2: Receiver Section

3.2. Flow of IBS



Step 1: All items will be at the mall equipped with RFID tags. When his phone placing a product in a wheelchair will identify the RFID reader to the processor.

Step 2: Second reading of the code on the ARM processor, after the game to the memory code, the processor reads the item name, price other information. Then, on the LCD screen. A language as the name, the price and the total amount of the object to the trailer available on the LCD screen.

Step 3: As stated above, the equipment, the costs for the total added. Moving of the carriage. At the same time, the information on the LCD screen. 16x2 LCD alphanumeric display mode. And even if we remove some, and press the Delete key to remove the object. The rate at which objects can be subtracted from the weight and to remove the entries on the LCD.

Step 4: LCD 4 is connected to the microcontroller 4bit mode. E 'is used for customers to learn about the measures to remove the client's action concerned an item, the price of the product and the total cost of the items in the cart.

Step 5: by check extended sites with respect to the transfer of data to a computer using ZigBee wireless transmitter in order to communicate with the processor. This is the RF

module 2.4GHz ISM band, which works, but not abandoned.

Step 6: ZigBee receiver connected PC using RS3232 sixth protocol, which receives your payment information and your computer for printing. This file contains all the information to obtain with respect to the total weight of the object. All prices are in Visual Basic 6.0. It shows the name of each product, the associated costs and the total weight of all the elements. The bill was introduced in Britain after the information, click on the screen.

Step 7: The information ZigBee all the details on the name of the item, which is the price, etc. Buy 8125 kHz RFID-tag by means of a passive type tag. A transponder (tag) at the object. RFID tag is a microchip and an antenna less. RFID tags come in various sizes, shapes and materials. The communication takes place between the RFID reader and the wireless tag, and often no line of sight between devices.

Step 8: RFID readers can read through everything. RFID reader transmits a low-power radio waves in the area is used to power the tag in order to transmit all the information on the chip.

Step 9: We will use ZigBee modules (transmitter and receiver) to pass a bill chariot your computer wireless to the table mode when the customer available products in the shopping cart and pay up to the counter. The people at the counter clicks to get the information and the data transfer from the wheelchair to a computer via ZigBee.

Step 10: LCD is used as the main output devices customers. It appears that the contents of the file, the price and the total weight, and other user information.

4. APPLICATIONS OF THE IBS & RESULTS

The main application of this system in the commercial centers to adjust the time and improving to reduce the quality of service of joy. It can be used everywhere and commercial markets where the bar codes are used for a good solution to the technology of the bar-code. It can also be used for the keeping in stock. The trolley is easy to use and requires no special training

Fig: 4.2 Kit at the Trolley side

Access cover wireless hardware module design skills, as shown below. There are two separate parts, the receiver and the Books section. ZigBee module for wireless data transfer. Figure 8. These models are created using Visual Basic. It turns out that the cash register computer. All costs and food seemed to be a number.

5. CONCLUSIONS

Finally, it should be emphasized that the inspiration and ideas to work after seeing a large number of payments and the fight against the sale Bazaar retail. When you work in this paper to learn about RFID technology, embedded systems and wireless systems, particularly in ZigBee modules savings system level in the supply chain. At the same time, but also the need for a sale. So it saves time and guarantees for each course. It is an automatic payment system for the purchase of the wind, and other useful products, such as the release of the workers, repetitive input, so that the working efficiency in the theft and supplies.

REFERENCES

[1] Stephen.B.Miles and Sanjay. E.Sharma, RFID Technology and Applications, Cambridge University Press, 1st edition.

[2] Myke Predko, Programming and Customi-sing the ARM7 Microcontroller, Tata Mc. graw- hill,

[3] http://www.scribd.com/doc/60264988/IEEE-

[4] Projects-in-Embedded-Systems

[5] http://www.engineersgarage.comirfid-radio-frequency-identification-and detection

[6] Suryaprasad, J; Kumar, B.O.P.; Roopa, D; Arjun A.k: Network Embedded systems for Enterprise Applications (NESEA), 2011 IEEE 2nd International conference.

[7] http. //truthdive com/2012/01/18/Now-a-shopping-cart-to-make-long-cash-countergueues-a-thing-of-the-past.html.

Fig: 4.1 Receiver Kit

Efficient Mining Of Top-K High Utility Itemsets From Uncertain Databases

Vamsinath Javangula [1], Suvarna Vani Koneru [2], Haritha Dasari[3]

[1] *Cse Department, P.B.R V I TS, Kavali, Andhra Pradesh, India*
vamsi.img@gmail.com

[2]*Cse Department, Velagapudi Ramakrishna Siddhartha Engineering College, Kanuru, Andhra Pradesh, India*
suvarnavanik@gmail.com

[3]*Cse Department, University College of Engineering College, Kakinada, JNTUK, Andhra Pradesh, India*
harithadasari9@yahoo.com

***Abstract*—**

Mining High-utility itemsets is vital issue in data mining research. Utility Mining have various Applications in cross-marketing retail stores, website click stream analysis and biomedical applications. The algorithms currently proposed for mining on data which defined exactly. Normally when big data received through various sensor networks due to noise presence it deviate from original values and data becomes uncertain. High utility itemsets are identified in uncertain data by setting proper minimum utility, minimum potential probability thresholds. Normally algorithms generate large number of high utility itemsets and lot time is consumed if proper threshold are not identified. In this paper, both of this problems addressed in an efficient framework to mine Top-K Uncertain High-Utility Itemsets (TKUHUI), is proposed, where k is the required high utility itemsets. Extensive experimental results on real and synthetic datasets show that TKUHUI is both efficient and scalable.

I. INTRODUCTION

Data Mining is popularly referred as information Discovery in knowledge (KDD). Data Mining projected varied techniques like Classification, Clustering, Association Rule Mining, Frequent Pattern Mining, and outlier analysis. In e-commerce, inventory management data processing play a significant role. In these areas frequent pattern mining terribly helpful.

Frequent pattern Mining is employed to catch the frequent patterns from transaction databases. Association rules mining (ARM) [2] model works based on the if an item is exist in the transaction or not by examining the all database items equally important. Frequent itemsets known by ARM will provide little profit among the overall profit,but non-frequent items can provide highest amount of profit..

In reality, a retail business could also be curious about distinguishing its most precious customers (customers who contribute a large amount of the profits to the company). These are the shoppers, who could purchase full priced things, high margin things, which can be absent from an most of the transactions as a result of most customers don't obtain these things. During a Association Rule Mining, these transactions representing extremely profitable customers could also be omitted. Utility mining is probably going to be helpful during a wide selection of sensible applications.

Recently, a utility mining model was outlined [2]. Utility could be a parameter of however "useful" an itemset is. The goal of utility mining is to spot high utility itemsets that drive an major portion of the total utility. Historical ARM drawback could be a special case of utility mining, wherever the utility of every item is often one and also the sales amount is either zero or one.

There is no efficient strategy to find all the high utility itemsets due to the nonexistence of "downward closure property" (anti-monotone property) in the utility mining model. A heuristics [2] is used to predict whether an itemset should be added to the candidate set

In MEU (Mining using Expected Utility) the prediction usually overestimates, particularly at the starting stages, wherever the no.of candidates approaches the quantity of all the mixtures of items. Such needs will simply overwhelm the memory area available and computation power of most of the machines. Additionally, MEU might miss some high utility itemsets once the variance of the itemset supports is more.

The challenge of utility mining is in limiting the scale of the candidate set and simplifying the computation for hard the utility. so as to tackle this

challenge, a Two-Phase algorithmic program to mine high utility itemsets.

Table 1. A transaction database

(a) Transaction table. Each row is a transaction. The columns represent the number of items in a particular transaction. TID is the transaction identification number

TID \ ITEM	A	B	C	D	E
T_1	0	0	18	0	1
T_2	0	6	0	1	1
T_3	2	0	1	0	1
T_4	1	0	0	1	1
T_5	0	0	4	0	2
T_6	1	1	0	0	0
T_7	0	10	0	1	1
T_8	3	0	25	3	1
T_9	1	1	0	0	0
T_{10}	0	6	2	0	2

(b) The utility table. The right column displays the profit of each item per unit in dollars

ITEM	PROFIT($)(per unit)
A	3
B	10
C	1
D	6
E	5

(c) Transaction utility (TU) of the transaction database

TID	TU	TID	TU
T_1	23	T_6	13
T_2	71	T_7	111
T_3	12	T_8	57
T_4	14	T_9	13
T_5	14	T_{10}	72

Our algorithm easily handles very large databases that existing algorithms cannot handle.

The rest of this paper is organized as follows. Section 2 overviews the related work. In Section 3, we detailed about background and problem definition. In section 4 we proposed our algorithm and techniques. Section 5 presents our experimental results and we summarize our work in section6.

II. Related Work

HUIM is totally new from FIM and ARM, considers local transaction utility (occur quantity) and external utility (unit profit) to discover itemsets from the quantitative databases which are profitable. The HUIM was proposed first by Chan et al. [8]. Yao et al. [21] then defined a peripheral unified framework for HUIM. Since the ARM downward closure property not work for HUIM, Liu et al. [15] designed the TWU model to maintain the transaction-weighted downward closure (TWDC) property, used to prune unpromising candidates for mining HUIs in a level-wise mechanism. Several mining HUIs such as IHUP [6], UP-growth [18], and UP-growth+ [19] based on tree structures have been extensively studied.

Based on these pattern-growth approaches, a lot of computations area unit still needed to come up with and keep the massive variety of discovered candidates for mining the real HUIs. To overcome the above drawbacks of previous HUIM, the HUI-Miner techniques [14] was used to directly mine HUIs to avoid the many database scans exclude candidate generation by the designed utility-list structure.

The FHM algorithm [17] was further proposed to enhance the performance of HUI-Miner by analyzing the co-occurrences among 2-itemsets. Instead of traditional HUIM, the variants of HUIM have been also extended and developed [12,20]. The development of other algorithms for HUIM is still in progress, but most of them are processed to handle precise data, the PHUIM framework [13] is the only work which focuses on mining high-utility itemsets on uncertain data.

The MUHUI algorithm[22] is proposed for in uncertain databases to find potential high-utility itemsets (PHUIs) using probability-utility-list (PU-list) structure, the MUHUI method directly mine PHUIs without n-itemsets generation and may cut back the development of PU-lists for various unimpressive itemsets by effective pruning plans, therefore greatly up the mining performance.

An TKU (Top-K Utility itemsets mining)[23] is projected for mining such itemsets while not setting min_util. many options were designed in TKU to unravel the new challenges raised during this downside,just like the absence of anti-monotone property and therefore the demand of lossless results. Moreover, TKU incorporates many novel ways for pruning the search area to attain high potency.

III. Background & Problem Definition

Researches that assign completely special weights to items introduced in [3, 4, 5, 6]. These weighted ARM models is new to utility mining. Itemset share practice is proposed in [7].It can be a utility as a result of it reflects the impact of the sales quantities of items on the price or profit of associate itemset. Many heuristics has been planned for utility Mining.

A utility mining algorithm program is proposed in [8], wherever the technique of “useful” is defined as a itemset that supports a particular objective that folks need to attain.

We start with the definition of a set of terms that leads to the formal definition of utility mining problem. The same terms are given in [2].

- $I = \{i_1, i_2, \ldots, i_m\}$ is a set of items
- $D = \{T_1, T_2, \ldots, T_n\}$ be a transaction database where each transaction Ti $\in$ D is a subset of *I*.
- $o(i_p, T_q)$, *local transaction utility value*, represents the quantity of item i_p in transaction T_q. For example, $o(A, T_8) = 3$, in Table 1(a)
- $u(i_p, T_q)$, *utility*, the quantitative measure of utility for item i_p in transaction T_q, is defined as $o(i_p,T_q) \times s(i_p)$. For example, $u(A, T_8) = 3 \times 3 = 9$, in Table 1
- $u(X, T_q)$, *utility of an itemset X in transaction* T_q, is defined as

$$\sum_{i_p \in X} u(i_p, T_q)$$

where $X = \{i_1, i_2, \ldots, i_k\}$ is a *k*-itemset, $X \subseteq T_q$ and $1 \leq k \leq m$.

- $u(X)$, *utility of an itemset X*, is defined as

$$\sum_{Tq \in D \ \wedge X \subseteq Tq} u(X, Tq)$$

Utility mining is to find all the itemsets whose utility values are beyond a user specified threshold. An itemset *X* is a *high utility itemset* if u(X) $\geq$ ε, where X $\subseteq$ I and ε is the minimum utility threshold, otherwise, it is a *low utility*

- *itemset*. For example, in Table 1, $u(\{A, D, E\}) = u(\{A, D, E\}, T_4) + u(\{A, D, E\}, T_8) = 14 + 32 = 46$. If ε =120, {A, D, E} is a low utility itemset
- The potential probability of an itemset *X* in *D* is denoted as *Pro*(*X*), which can be defined as: $Pro(X) = \sum_{X \subseteq T_q \wedge T_q \in D} p(X, T_q)$.
- An itemset *X* in an uncertain database *D* is defined as a potential high-utility itemset (PHUI) if it satisfies the following two conditions:
  (1) *X* is a HUI w.r.t. $u(X) \geq \varepsilon \times TU$;
  (2) $Pro(X) \geq \mu \times |D|$.
- A desired TKUHUI indicates the itemset has both high potential probability and high utility value

Problem Statement: Given an uncertain database *D* with total utility is *TU*, the minimum utility threshold and the minimum potential probability threshold are respectively set as ε and μ. The problem of potential high-utility itemset mining (TKUHUI) from uncertain data is to mine TOP-K UHUIs.

IV. THE UNCERTAIN TOP-K UTILITY ITEMSET ALGORITHM

Although the PHUI-List algorithm has better performance compared to the upper-bound-based PHUI-UP algorithm [13], however, it explores the search space of itemsets by generating itemsets, and a costly join operation of probability-utility-list (PU-list) has to be performed recursively to evaluate the probability and utility information of each itemset. By utilizing the PU-list structure, a more efficient TKUHUI algorithm is proposed here to improve the performance

A. The PU-list Structure

The PU-list structure [13] is a new vertical data structure; it incorporates the probability and utility properties to keep necessary information from uncertain data in terms of TID information, probability, utility, and remaining utility information.

Let an itemset X and a transaction (or itemset) T such that X $\subseteq$ T,the set of all items from T that are not in X is denoted as T\X, and the set of all the items appearing after X in T is denoted as T/X. Thus, T/X $\subseteq$ T\X.For example, consider X = {CD} and transaction T7 in Table 1, T7\X = {AE}, and T7/X = {E}.

The PU-list of an itemset X in a database is denoted as X.PUL. It contains an entry (element) for each transaction Tq where X appears (X $\in$ Tq $\subseteq$ D). An element consists of four fields: (1) the tid of X in Tq (X$\subseteq$Tq$\in$D); (2) the probabilities of X in Tq (prob); (3) the utilities of X in Tq (iu); and (4) the remaining utilities of X inTq (ru), in which ru is defined as $X.ru(Tq)=\sum_{i_j \in (T_q/X)} u(i_j, T_q)$.

Therefore, all necessary information from uncertain data can be compressed into the designed PU-list structure without losing any useful information. Thanks to the property of PU-list, the probability and utility information of the longer *k*-itemset can be built by joining its parent node and uncle node, i.e., (*k*-1)-itemset. The join operation can be easily done without rescanning the database.

The construction procedure of the PU-list is recursively processed if it is necessary to determine the k-itemsets in the search space, details of the construction can be referred to [13]. Note that it is necessary to initially construct the PU-list of the complete set of HTWPUI1 [13] as the input for the later recursive process.The PU-list is constructed in TWU ascending order as (B ←A ← D ← E ← C),

which is shown in Fig. 1.

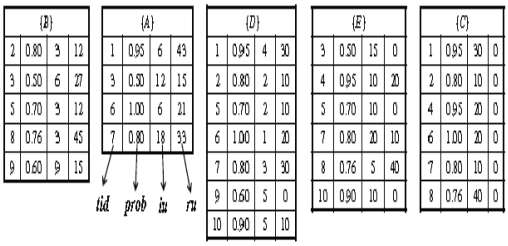


Fig. 1. Constructed PU-list structure of HTWPUI[1].

The sum of the utilities and remaining utilities of an itemset X in D, denoted as X.IU and X.RU, respectively, which can be defined as:

$$X.IU = \sum_{X \subseteq T_q \wedge T_q \in D} (X.iu), X.RU = \sum_{X \subseteq T_q \wedge T_q \in D} (X.ru).$$

B.Search Space and Properties

Based on the PU-list structure, the search space of the propose TKUHUI algorithm can be represented as the Set-enumeration tree by the TWU values of the 1-items in the set of HTWPUI1 in ascending order, as shown in Fig. 2 (left).Based on the constructed Set-enumeration tree, the following lemmas can be obtained.

Lemma 1. *The sum of all the probabilities of any node in the Set-enumeration tree is greater than or equal to the sum of all the probabilities of any of its child nodes.*

Proof. Assume a (k-1)-itemset w.r.t. a node in the Set-enumeration tree be Xk−1(k ≤ 2), and any of its child nodes be denoted as Xk. Since p(Xk, Tq) =p(Tq) for any transaction Tq in D, it can be found that: p(Xk,Tq)/p(Xk−1,Tq) = p(Tq)/p(Tq) = 1.

Since Xk−1 is subset of Xk, the TIDs of Xk is the subset of the TIDs of Xk−1,thus,

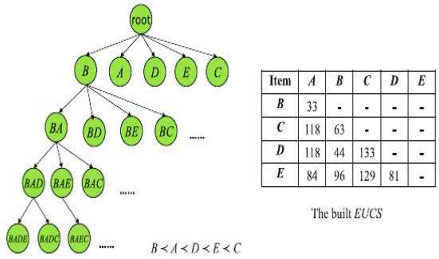
$$Pro(X^k) = \sum_{X^k \subseteq T_q \wedge T_q \in D} p(X^k, T_q) \leq \sum_{X^{k-1} \subseteq T_q \wedge T_q \in D} p(X^{k-1}, T_q) = Pro(X^{k-1}).$$

Lemma 2. For any node X in the Set-enumeration tree, the sum of X.IU andX.RU is greater than or equal to the sum of all the utilities of any one of its child nodes.

Proof. From [13], this lemma holds.

C.Proposed Pruning Strategies

The proposed algorithm uses an internal variable named border minimum utility threshold (denoted as border_min_util) which is initially set to 0 and raised dynamically after a sufficient number of itemsets with higher utilities has been captured during the generation of TKUHUIs. The development of the proposed method is based on the following definitions and lemmas.



Item	A	B	C	D	E
B	33	-	-	-	-
C	118	63	-	-	-
D	118	44	133	-	-
E	84	96	129	81	-

The built *EUCS*

Fig. 2. Constructed Set-enumeration tree and *EUCS*.

Lemma 1. Let P= be a set of itemsets (m ≥ k), where Xi is the i-th itemset in P and u(Xi) ≥ u(Xj), ∀ i < j. (In other words, Xi is the itemset with the i-th highest utility in P). For any itemset Y, if u(Y) < u(Xk), Y is not a top-k high utility itemset.

Rationale. According to Definition 10, if there exist k itemsets whose utilities are higher than the utility of Y, Y is not a top-k high utility itemset.

Lemma 2. Let P= be a set of itemsets (m ≥ k), where Xi is the i-th itemset in P and u(Xi) ≥ u(Xj), ∀ i < j. If δP = u(Xk), fH(D, δ*)⊆ fH(D, δP). Rationale. Let H be the complete set of top-k high utility itemsets. If |H| ≥ k, δ* = min{u(X)| X∈H} (by Definition 11). Because δ* = min{u(X)| X∈H} ≥ min{u(Xi)| Xi ∈ P, 1 ≤ i ≤ k} = u(Xk) = δP, δ* ≥ δP and fH(D, δ*)⊆ fH(D, δP).

Example 3. Suppose k = 4 and border_min_util = 0 initially. Let P be the set of 1-items in D. Then P = {{A}:20, {D}:20, {B}:16, {E}:15, {C}:13, {G}:7, {F}:5}, where the number beside each item is its exact utility. By Lemma 1, items {C}, {G}, {F} are unpromising to be the top-4 high utility itemsets. Therefore border_min_util can be raised to 15, the 4th highest utility value in P, and no top-k high utility itemset will be missed.

After raising border_min_util, the algorithm performs the UPGrowth search procedure with min_util = border_min_util to generate TKUHUIs. Although Lemma 1 provides a way to raise border_min_util, it cannot be applied during the generation of TKUHUIs in phase I. This is because the exact utilities of the TKUHUIs are unknown during phase I. One of the solutions to this problem is to use lower bound of the utility of TKUHUI to raise the border_min_util. A lower bound of the utility of an itemset can be estimated by the following definitions.

Lemma 3. Let C = be a set of itemsets (m ≥ k), where Xi is the i-th itemset in C and MIU(Xi) ≥ MIU(Xj), ∀ i < j. For any itemset Y, if TWU(Y) < δC = min{MIU(Xi) | Xi ∈ C, 1 ≤ i ≤ k}, Y is not a top-k high utility itemset. Rationale. According to Definition 8, u(Y) ≤ TWU(Y). If TWU(Y) < δC, u(Y) < δC. Besides, u(Y) < MIU(Xi) ≤ u(Xi), Xi ∈ C, 1 ≤ i ≤ k. According to Definition 10, if there

exist k itemsets whose utilities are higher than the utility of Y, Y is not a top-k high utility itemset.

Lemma 4. Let C = be a set of itemsets ($m \geq k$), where Xi is the i-th itemset in C and MIU(Xi) ≥ MIU(Xj), $\forall$ i < j. If δC = MIU(Xk), fH(D, δ^*) $\subseteq$ fH(D, δC). Rationale. Let H be the complete set of top-k high utility itemsets. If |H| ≥ k, δ^* = min{u(X)| X$\in$H} (by Definition 10). Because δ^* = min{u(X)| X$\in$H} ≥ min{u(Xi)| Xi $\in$ C, $1 \leq i \leq k$} ≥ min{MIU(Xi) | Xi $\in$ C, $1 \leq i \leq k$}= MIU(Xk), we have $\delta^* \geq \delta C$ and fH(D, δ^*) $\subseteq$ fH(D, δC).

Lemma 5. For any itemset X, if TWU(X) < border_min_util ≤ δ^* , X and all its supersets are not top-k high utility itemsets.

Based on the above lemmas and definitions, we have the following ideas to raise border_min_util during the generation of TKUHUIs. As soon as a candidate X is found by the UP-Growth search procedure, we check whether its estimated utility (i.e, TWU(X)) is higher than border_min_util. If TWU(X) < border_min_util, X and all its supersets are not top-k high utility itemsets (Lemma 5). Otherwise, we check whether its MAU is higher than border_min_util. If MAU(X) < border_min_util, X is not a top-k high utility itemset (Lemma 6). Otherwise, X is considered as a candidate for phase II and it is outputted with its estimated utility value according to Lemma 7. If X is a valid TKUHUI and MIU(X) ≥ border_min_util, MIU(X) can be used to raise the border_min_util (Lemma 3). To efficiently update border_min_util, we use a min-heap structure L to maintain the k highest MIUs of the TKUHUIs until now. Once k MIUs are found, border_min_util is raised to the k-th MIU in L according to Lemma 3. Each time a TKUHUI X is found and its MIU is higher than border_min_util, X is added into L and the lowest MIU in L is removed. After that, border_min_util is raised to the k-th MIU in L. The algorithm continues searching for more TKUHUIs until no candidate is found by the UP-Growth search procedure. Figure 3 gives the pseudo code for the above processes.

```
If(TWU(X) ≥ border_min_util and MAU(X) ≥ border_min_util)
 {
Output X and min{TWU(X), MAU(X)}
If (MIU(X) ≥ border_min_util) {
Add X to L and raise border_min_util by MIU(X)}
}
else {
 X is not a valid TKUHUI
 }
```

Fig 3. The pseudo code for the strategy MC

V. Experimental Evaluation

All the algorithms are implemented in C++. Experiments are performed on a computer with 2.93 GHz Intel Core 2 Processor and 4 GB memory. The operating system is Ubuntu 12.04.d.All of the algorithms are implemented in Java. Different types of real world datasets were used in the experiments. Foodmart, a sparse dataset, was acquired from Microsoft foodmart 2000 database [1]; Mushroom, a dense dataset, was obtained from the FIMI Repository [1]; Chainstore, a large dataset, was obtained from NU-MineBench 2.0 [15]. The two datasets Foodmart and Chainstore already contain unit profits and purchased quantities. For Mushroom dataset, unit profits for items are generated between 1 and 1000 by using a log-normal distribution and quantities of items are generated randomly between 1 and 5, as the settings of [19]. Table 2 shows the characteristics of the datasets used in the experiments.

Dataset	#Transactions	Avg. length	#Items	Type
Foodmart	4,141	4.4	1,559	Sparse
Mushroom	8,124	23.0	119	Dense
Chainstore	1,112, 949	7.2	46,086	Sparse Large

Experiments are compared under varied minimum utility thresholds (abbreviated as MUs) with the fixed minimum potential probability threshold (abbreviated as MP). The runtime results under varied MUs with a fixed MP are shown in Fig. 4.

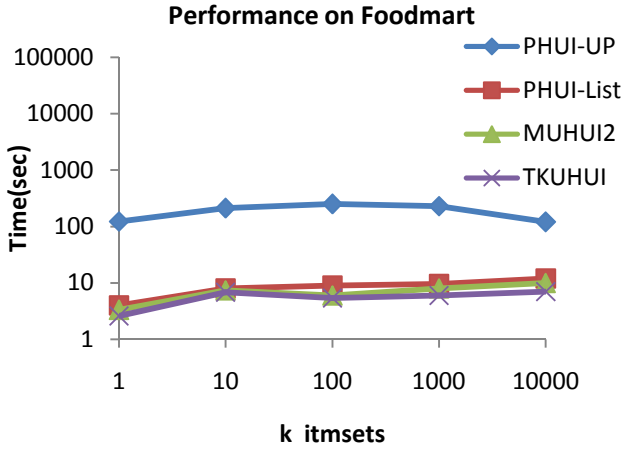


Fig 4. Performance of algorithms on Foodmart

From Fig.4, it can be observed that the runtime of all the algorithms is decreased along with the increasing of MU. In particular, the proposed TKUHUI algorithm is generally up to almost one or two orders of magnitude faster than the PHUI-UP algorithm, and also outperforms the state-of-the-art PHUIList algorithm on all datasets. It is reasonable since the upper-bound-based generate-and-test mechanism has worse results than the vertical PU-list-based approaches. Besides, the TKUHUI algorithm uses pruning strategies to early prune unpromising itemsets and search space, which can avoid the costly join operations of a

huge number of PU-lists for mining PHUIs, but the PHUI-List. When the MU is set quite low, longer patterns of HTWPUIs are first discovered by the PHUI-UP algorithm, and thus more computations are needed to process with the generate-and-test mechanism, especially in a dense dataset. from the Set-enumeration tree without candidate generation in a level-wise way,it can effectively avoid the time-consuming dataset scan.

VI. CONCLUSION

In this paper, we have proposed an efficient algorithm named TKUHUI for mining top-k high utility itemsets from transaction databases. TKUHUI guarantees there is no pattern missing during the mining process. We develop strategies for to raise the border minimum utility threshold and reduce the search space and number of generated candidates. Moreover, a strategy is designed decrease the number of checked candidates. The mining performance is enhanced significantly since both the search space and the number of candidates are effectively reduced by the proposed strategies. In the experiments, different types of real datasets are used to evaluate the performance of our algorithm. The experimental results show that TKUHUI outperforms the baseline

References

1. Frequent itemset mining dataset repository. http://fimi.ua.ac.be/data/
2. Aggarwal, C.C.: Managing and mining uncertain Data (2010)
3.Aggarwal, C.C., Yu, P.S.: A survey of uncertain data algorithms and applications.IEEE Trans. Knowl. Data Eng. 21(5), 609–623 (2009)
4. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: The International Conference on Very Large Data Bases, pp. 487–499 (1994)
5. Agrawal, R., Srikant, R.: Quest synthetic data generator. http://www.Almaden.ibm.com/cs/quest/syndata.html
6. Ahmed, C.F., Tanbeer, S.K., Jeong, B.S., Le, Y.K.: Efficient tree structures for high utility pattern mining in incremental databases. IEEE Trans. Knowl. Data Eng. 21(12), 1708–1721 (2009)
7. Bernecker, T., Kriegel, H.P., Renz, M., Verhein, F., Zuefl, A.: Probabilistic frequent itemset mining in uncertain databases. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 119–128 (2009)
8.Chan, R., Yang, Q., Shen, Y.D.: Mining high utility itemsets. In: IEEE International Conference on Data Mining, pp. 19–26 (2003)
9. Chui, C.-K., Kao, B., Hung, E.: Mining frequent itemsets from uncertain data. In:Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp.47–58. Springer, Heidelberg (2007)
10. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: a survey. ACM Comput. Surv. 38 (2006)
11. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min. Knowl. Disc. 8(1), 53–87(2004)
12. Lin, J.C.W., Gan, W., Hong, T.P., Tseng, V.S.: Efficient algorithms for mining up-to-date high-utility patterns. Adv. Eng. Inform. 29(3), 648–661 (2015)
13. Lin, J.C.W., Gan, W., Fournier-Viger, P., Hong, T.P., Tseng, V.S.: Mining potential high-utility itemsets over uncertain databases. In: ACM ASE BigData & Social Informatics, p. 25 (2015)
14. Liu, M., Qu, J.: Mining high utility itemsets without candidate generation. In: ACM International Conference on Information and Knowledge Management, pp.55–64 (2012)
15. Liu, Y., Liao, W., Choudhary, A.K.: A two-phase algorithm for fast discovery of high utility itemsets. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 689–695. Springer, Heidelberg (2005)
16. Microsoft: Example Database foodmart of Microsoft Analysis Services.http://msdn.microsoft.com/en-us/library/aa217032 (SQL . 80) . aspx
17. Fournier-Viger, P., Wu, C.-W., Zida, S., Tseng, V.S.: FHM: faster high-utility itemset mining using estimated utility co-occurrence pruning. In: Andreasen, T., Christiansen, H., Cubero, J.-C., Ra´s, Z.W. (eds.) ISMIS 2014. LNCS, vol. 8502,pp. 83–92. Springer, Heidelberg (2014)
18. Tseng, V.S., Wu, C.W., Shie, B.E., Yu, P.S.: UP-growth: an efficient algorithm for high utility itemset mining. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 253–262 (2010)
19. Tseng, V.S., Shie, B.E., Wu, C.W., Yu, P.S.: Efficient algorithms for mining high utility itemsets from transactional databases. IEEE Trans. Knowl. Data Eng.25(8), 1772–1786 (2013)
20. Wu, C.W., Shie, B.E., Tseng, V.S., Yu, P.S.: Mining top-k high utility itemsets. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 78–86 (2012)
21. Yao, H., Hamilton, H.J., Butz, C.J.: A foundational approach to mining itemset utilities from databases. In: SIAM International Conference on Data Mining, pp.211–225 (2004)
22. Jerry Chun-Wei Lin, Wensheng Gan, Philippe Fournier-Viger,Tzung-Pei Hong, Vincent S. Tseng.: Efficient Mining of Uncertain Data for High-Utility Itemsets. In: B. Cui et al. (Eds.): WAIM 2016, Part I, LNCS 9658, pp. 17–30. Springer, Switzerland (2016)
23. Cheng Wei Wu, Bai-En Shie, Philip S. Yu, Vincent S. Tseng.: Mining Top-K High Utility Itemsets. In: KDD'12, August 12–16, 2012, Beijing, China.

Enhanced Caesar Cipher algorithm with variable length key and increased cipher complexity

P. Srinivasa Rao
Research Scholar
Acharya Nagarjuna University
Nagarjuna nagar, Guntur
Andhra Pradesh, India.
Srinivas.pokuri2@gmail.com

Dr. D. Nagaraju
Prof. & Head, Dept. of Information Technology
Lakireddy Bali Reddy College of Engineering
L.B. Reddy Nagar, Mylavaram(M),
Andhra Pradesh, India.
dnagaraj_dnr@yahoo.co.in

***Abstract* — For every organization information is one of the most important assets. We put many efforts to protect systems and networks to achieve confidentiality, integrity and authentication. Its is necessary to implement a secured wrapper around the data, that is none other than encryption. "Encryption is the process of transforming plain text into the cipher text where plain text is the input to the encryption process and cipher text is the output of the encryption process". The process of decryption is transforming cipher text back into the plain text and this plain should be the same text which was sen by the sender before encryption. In this technological world cyber security still remains a hot topic in research because of the importance of information security. There is a huge demand for an algorithm to encrypt the data which take nearly infinity time to brute force with a single machine. Character substitution algorithms were the first encryption algorithms that came into existence. Even though they are considered as less secured, there were still research going on to increase it complexities in various aspects. We propose the following approach to achieve a strong character substitution algorithm with less time complexity when compared to other algorithms providing the same level of security covering all the required outcomes of encryption.**

1. INTRODUCTION

The word Cryptography is derived from two *Greek* words '*K*ryptos*'* and '*Graphy'* which means '*Secret' and 'Writing'* respectively. Cryptography is the study and practice of secret writing and its techniques. We can simply say it is an art of hiding information or transforming information into raw data/unreadable data. From the beginning we have used many types of algorithms for a secure data transmission.[1] A cryptographic algorithm is combination of mathematical functions and some predefined set of steps to perform encryption and decryption of the data. Some times this algorithm can be combination of two or more existing algorithms. On an whole the main objective is to make it as strenuous as possible to get plain text back from the generated cipher text without using the symmetric key used by the sender as a part of encryption.[2] If we a good encryption algorithm, then there will be no technique will be better than applying the brute force method of trying the every possible combination of key. Mode of encryption has changed many times, some think character replacement ciphers, others block ciphers, some goes for public and private keys. So there are many types of encryption, we select the mode of encryption based on the type of data we're going to secure. Not only type, it has multiple variants to decide which algorithm should have to select like time complexity, reliability of the mode.

We would like to present an algorithm to encrypt and decrypt the data which is based on ASCII values of characters in the plain text. In this algorithm ASCII values are used to encrypt data. Both the time and cipher complexity is completely dependent on the user given password or key. We're going to take a variable length key as input and after that we process this user given key and generates a suitable key for the given plain text. With this we increased the complexity of guessing password.

The use given key is modified and the algorithm takes this as key and the plain text as input and generates cipher text in form of ASCII values. Again these ASCII values can be casted back to the plain text with the same key given by the user.

2. LITERATURE SURVEY

2.1 Substitution Techniques:

A substitution technique is one in which the letters of plain text are replaced by other letters or by numbers or symbols. If the plain text is viewed as a sequence of bits, then substitution involves replacing plain text bit patterns with cipher text bit patterns.

2.1.1 Caesar's Shift Cipher:

From the beginning of 1900 BCE so many tried to develop best ciphering technique but so many of them were pointed with a breach. Every technique has its own advantages and disadvantages. "Once Gaius Julius Caesar has manipulated his message with 3 character shifting right to transfer message to his team, that was a substitution cipher concept with symmetric key encryption". This algorithm was used for some duration and later that was considered very weak as its key complexity is very low. But considering this as a base so many reliably strong algorithms were developed.

2.1.1.1 Methodology :

In this algorithm the process of both Encryption(E) and Decryption(D) is done with the same secret key(K). Each character in the plain text will be replaced by a value which is equal to ***K+valueOfCharacter modulo 26*** number of characters to right, i.e., let K =10 then 'a' is replaced by 'k' , 'b' by l, … 'z' by 'j'. Let C be the sequence value/ positional value of that and K be the secret key value.

For *Encryption* := (C+K) ***mod*** 26
For *Decryption* := (C-K) ***mod*** 26

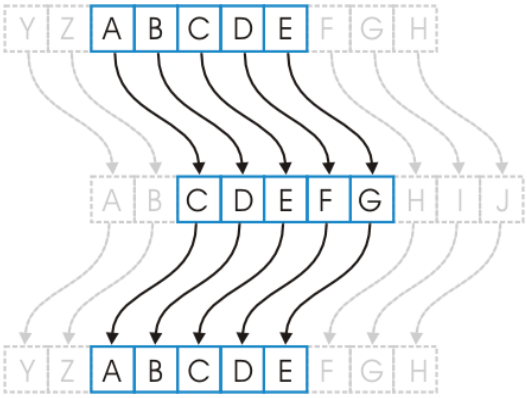


Fig -2.1 *Shifting of Characters*

2.1.1.2 Drawback :

Abiding this algorithm, the characters are shifted in a circular manner. If we try a brute force method of rotating values in a loop ranging [0,26) we can get the original plain text very easily and this means the key combination complexity of this algorithm is very poor.

2.2 Feistel Network :

"A Feistel network is a general method of transforming any function (usually called an F-function) into a permutation. It was invented by Horst Feistel and has been used in many block cipher designs."

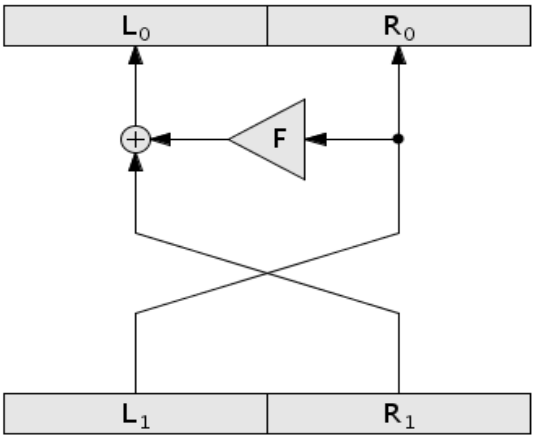


Fig- 2.2 *Feistel Network Working*

2.2.1 Working of Feistel Network :

The above figure (*Fig-2.1*) represents the following steps.

Step1: Split each block into halves,
Step2: Right half becomes new left half.
Step3: New right half is the final result when the left half is XOR'd with the result of applying ***F*** to the right half and the key.

Note that previous rounds can be derived even if the function F is not invertible.

2.3 GOST :

GOST is an acronym for ***go***sudarstvennyy ***st***andart, which means State Standard in Russian. Its is a symmetric block cipher, which conforms to Feistel scheme. 64-bit blocks of data are submitted to the input and converted into 64-bit blocks of encrypted data by 256-bit key.

2.3.1 Working of GOST:

"In each round the right side of plain text messages is processed by function F, which converts data with three cryptographic operations: adding data and sub-key modulo 232, substitution of data using S-boxes, and left cyclic shift by 11 positions(we can see it clearly in the *Fig- 2.3*). Output of F-function is added modulo 2 to the left part of the plain text, then right and left sides are swapped for next round. The algorithm has 32 rounds. In the last round of encryption right and left parts are not swapped."

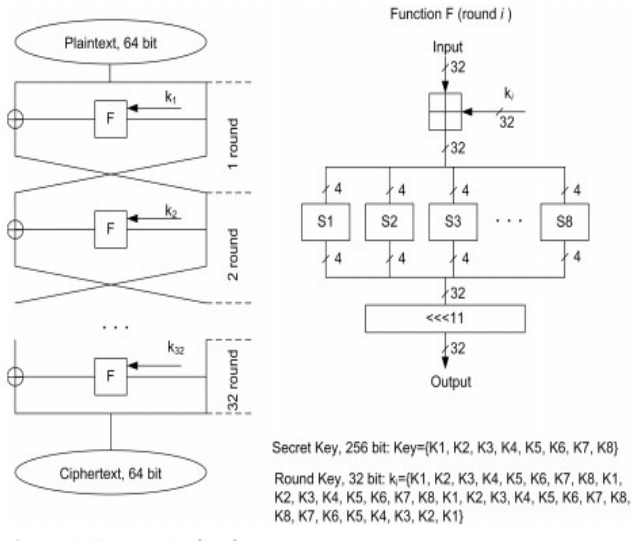


Fig- 2.3 *GOST encryption working*

2.4 Blowfish :

Blowfish was designed in 1993 by Bruce Schneier as a fast, free alternative to existing encryption algorithms. Blowfish is an unpatented and license-free, and is available free for all uses. It is a symmetric block cipher that can be effectively used for encryption and safeguarding of data. It takes a variable-length key, from 32 bits to 448 bits, making it ideal for securing data.

2.4.1 Blowfish working

There are five sub-key arrays: One 18-entry P-array (denoted as K in the *Fig 2.3* , to avoid confusion with the Plain text) and four 256-entry S-boxes (S0, S1, S2 and S3).

"Every round r consists of 4 actions: First, XOR the left half (L) of the data with the r^{th} P-array entry, second, use the XOR'ed data as input for Blowfish's F-function, third, XOR the F-function's output with the right half (R) of the data, and last, swap L and R."

The below figure *(Fig 2.4*) clearly states the "F-function splits the 32-bit input into four eight-bit quarters, and uses the quarters as input to the S-boxes. The S-boxes accept 8-bit input and produce 32-bit output. The outputs are added modulo 232 and XOR'ed to produce the final 32-bit output. After the 16th round, undo the last swap, and XOR L with K18 and R with K17 for output."

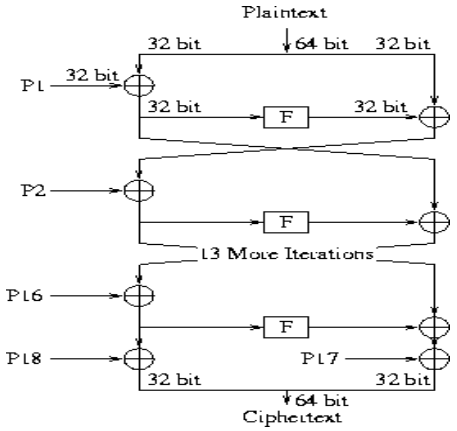


Fig- 2.4 *Blowfish Encryption working*

3. PROPOSED ALGORITHM

3.1 Encryption algorithm :

- Let user given text be **T** and its length be $\mathbf{T_L}$
- User given password be **P** and its length is $\mathbf{P_L}$
- We need to divide password for each part of the text to encrypt.
- Before that we need to check whether the password can exactly distributed to all the partitions of plain text.
- Partition length is ***part_len = TL /PL***
- If the password cannot be distributed evenly to all the partitions keep appending the mid value in the password array.
- Checking condition ***part_len*PL = TL*** if it can be distributed otherwise append value after every updated state until the condition satisfies.
- In this encryption the rounds of encryption is based on the updated password length $\boldsymbol{P_L}$

Now we generate keys from the password

- For $\boldsymbol{i_{th}}$ round of encryption the keys are generated from the password array ***P***
  - Calculate ***password_rotate*** = $(2*P_i+1) \bmod (8*P_L)$
  - Convert ***P*** array into bits string
  - Rotate ***password_rotate*** number of bits to left and cast back to integer array
  - Updated this keys in password and save this key list in a multi-dimensional array
- Like above we generate $\boldsymbol{P_L}$ number of key lists and saved for processing Encryption

Now Encryption method

- For $\boldsymbol{i_{th}}$ round of encryption, we have $\boldsymbol{i_{th}}$ list of keys generated
  - Calculate ***text_rotate*** = $(2*P_i+1) \bmod (8*T_L)$
  - Convert text ***T*** into bits string
  - Rotate ***text_rotate*** number of bits to left and cast back to integer array
  - Divide ***T*** into partitions of length ***part_len*** and assign each key in the list to the partition respectively.
  - Now **XOR** the values in the partition with the key assigned and store back
- Like above, we complete $\boldsymbol{P_L}$ number of rounds encryption

3.2 Decryption algorithm :

- Every step in decryption is same as in the encryption but we need to process two things in the reverse order
  - Key list to be reversed
  - Decryption process
    - First XOR the partitions with key values
    - Rotate as per the ***text_rotate*** value calculated

Working model of this algorithm is explained in the *Fig-3.1*. Here the number of partitions are depends on length of the user given password and also depends on the even distribution of keys to all partitions.

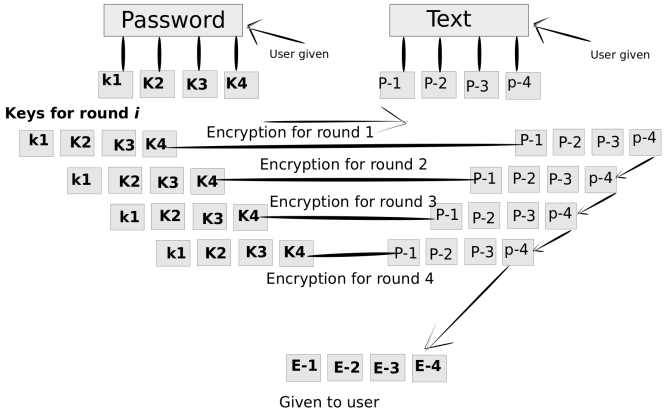


Fig- 3.1 *encryption in proposed algorithm*

4. IMPLEMENTATION

We have developed this algorithm in Python programming language and tested on various inputs. Let us take an example input and password to understand the process clearly

4.1 Reading Input

1. User given text ***T*** be "abak" and password ***P*** be "LDA"
2. It is clear that text length $T_L = 4$ and password length $P_L=3$
3. Calculate partition length $part\_len = T_L/P_L = 4/3 = 1$
   1. while $part\_len*P_L != T_L$
   2. add mid value if ***P*** array
   3. increment P_L
4. Now constraint of even distribution of keys is satisfied.
5. ASCII values of text ***T*** and password ***P*** are [98,97,98,107] and [76,68,65,68]

4.2 Generating keys

1. For round ***i=1***, ***password_rotate = (2*98+1)mod(8*4)***

Rotate ***password_rotate*** number of bits to left in this bit string and cast back to integers. The resultant array will be [152, 136, 130, 136]

2. For round ***i=2***, ***password_rotate = (2*136+1)mod(8*4)*** and the resultant key list is [49, 17, 5, 17]

3. Likewise keys for round ***3*** and ***4*** are [136, 136, 40, 137] and [68, 65, 68, 76] respectively

4.3 Encryption

1. The number of rounds of encryption is dependent on the length of the password updated as mentioned in the above algorithm.
2. The updated P_L=4, therefore round count of encryption of this text with respect to the updated key is 4.
3. For our understanding and convenience, we write text and key values in the number format.
4. For ***round-1***, the processing text will be [97, 98, 97, 107] and the key list used to encrypt this text is [152, 136, 130, 136], the first list formed .
5. As the ***round_number*** is ***1***, we use K_1 = 152 as the ***text_rotate*** value and $(2*K_1+1)mod(8*T_L)$ = 17, we rotate this bit string chunk 17 bits to left and cast back to integers.
6. Now the text values will become [194, 214, 194, 196], now masking of this text values is completed.
7. We process this text for XOR operations with keys [152, 136, 130, 136] with [194, 214, 194, 196] respectively will gives the result [90, 94, 64, 76]
8. Now the ***round-1*** is completed and after this round one the text values will be the input text values (***[90, 94, 64, 76])*** to the next round.
9. After ***round-2, round-3*** and ***round-4,*** the values of the text are transformed to [227, 227, 7, 115], [134, 111, 239, 79], [219, 77, 155, 146] respectively.
10. The values after the ***round-4 (or) round-n*** are given as cipher text to the other user.
11. The list [219, 77, 155, 146] is written in output file with space as delimiter.

- The process of encryption is explained step by step in the *Fig-3.1*

4.4 Decryption

1. The decryption process is just reverse to the *encryption*
2. Input to the decryption is the values that we have written in file after *encryption* process ie., [219, 77, 155, 146]
3. The process of ***key_generation*** is same as in the *encryption* process.
4. Here the ***key_list*** is reversed for the processing and all the input evaluation will be same as in the *encryption* process *(as mentioned above)*
5. The output of this encryption will be [98,97,98,107]

4.5 Encryption samples

1. Sample data is given input to the encryption algorithm developed in Python language.
2. Values are mentioned in the ASCII format for better understanding.

Values are given detailed in *Table-1* below

Input	***User key***	***Updated key***	***Output***
97, 98, 97, 107	76, 68, 65	76, 68, 65, 68	219, 77, 155, 146
• First test case of plain and cipher values			
97, 98, 97, 107	97, 98, 97, 107	97, 98, 97, 107	34, 3, 13, 247
• To observe the frequency difference for the change in key			
97, 97, 97, 97	97, 98, 99, 100	97, 98, 99, 100	79, 48, 125, 70
• Even if the plain text is same, the cipher text frequency changes with respect to the user given password.			

97, 98, 99, 100	97, 98, 99, 100	97, 98, 99, 100	79, 0, 93, 22
• If text and key are same, there is change in cipher with respect to frequency			

Table-1 *Sample input and outputs*

- This complexity much differs from the traditional cipher systems and can compete with the advanced encryption techniques which are having same time complexity.

5. TESTING

5.1 Time complexity

Here the time complexity is a variable and is dependent on the P_L, the length of the password given by user. Because the number of rounds is dependent on the user and if needs complexity he can give a key having greater length.

The operations performed in the algorithms are circular shifts, and these are easy to perform when they are considered as bit stings.

XOR operation is the only operation we use here and it is operating in low level. So it will not take much time when compared to other power operators.

5.2 Avalanche Effect

A minute change in plain text or key will give a great difference in the cipher text is called *Avalanche effect.* As mentioned in the above table (*Table-1*), changes in plain text and key are mentioned and we can observe the difference in the frequency of the cipher text.

Let we test for the reverse case of tampering the cipher text and its detection (mentioned in *Table 5.1)*.

Actual Cipher	***Tampered Cipher***	***Key***	***Actual Plain Text***	***Tampered Output***
219, 77, 155, 146	***219, 155, 146***	76, 68, 65	97, 98, 97, 107	***205, 134, 143***
• Here some bits are removed(8 bits) and the resultant plain text has no match to the expected plain text				
219, 77, 155, 146	***219, 77, 155, 146,146***	76,68,65	97, 98, 97, 107	***22, 161, 208, 17, 35***
• Here some bits are added (8 bits) and the resultant plain text has no match to the expected plain text				

Table 5.1 *Decryption outputs for tampered input*

5.3 Strengths of this algorithm

1. Enhanced the complexity of cipher text with reduced computational power.
2. Core concepts and problems of Caesar Cipher are focused and solved here.
3. Noval from the traditional methods.

6. FUTURE WORKS

We are testing this algorithm on very large and real time inputs like database, on-site encryption. For that we are optimizing and stabilizing this algorithm.

7. CONCLUSION

Based on the Caesar cipher technique this algorithm has introduced an innovative approach for character replacement encryption algorithms. Many existing algorithms have weaknesses caused by time delay and decreased security levels with poor design. This algorithm is tested with different cyber attacks and resulted to be secured still. Therefore, this algorithm can be a good alternative to other techniques in some applicable areas.

REFERENCES

[1]http://www.iaeng.org/publication/WCECS2012/WCECS2012_pp979-982.pdf

[2] http://www.enggjournals.com/ijcse/doc/IJCSE12-04-09-103.pdf

[3] https://www.cs.rit.edu/~ark/fall2013/462/module03/fig3.png

[4]

Gary C. Kessler “An Overview of Cryptography” , May 1998

http://www.garykessler.net/library/crypto.html

[5] McGraw, Gary, Felten, Edward F, Securing Java, New York: New York, John Wiley & Sons, 1999

[6] Feghhi, Jalal, Feghhi, Jalil, Williams, Peter, Digital Certificates, Addison Wesley Longman, Inc., 1999

[7]Christoyannis, Costas, http://www.hack.gr/users/dij/crypto/

[8] SSH Communications Security,

http://www.ssh.fi/tech/crypto/intro.html#algorithms

[9] http://practicalcryptography.com/ciphers/caesar-cipher/

[10] http://www.securingthehuman.org/newsletters/ouch/issues/OUCH-201107_en.pdf

[11]http://msdn.microsoft.com/en-us/library/windows/desktop/aa381939%28v=vs.85%29.aspx

[12]http://etc.usf.edu/clipart/87800/87831/87831_cipher.htm

[13] Handbook of Applied Cryptography

[14] http://www.apprendre-en-ligne.net/crypto/bibliotheque/PDF/Kwang.pdf

[15] http://data-informed.com/issues-address-evaluating-data-encryption-cloud/

[16] https://www.wikiwand.com/simple/Avalanche_effect

[17] https://www.schneier.com/academic/blowfish/

[18]http://www.cs.trincoll.edu/~crypto/historical/caesar.html

[19] http://www.gfi.com/blog/security-101-encryption-terminology/

[20] http://csrc.nist.gov/groups/STM/cavp/

[21]https://www.owasp.org/index.php/Cryptographic_Storage_Cheat_Sheet#Rule_-_Use_approved_cryptographic_modes

[22]https://www.owasp.org/index.php/Guide_to_Cryptography#How_to_determine_if_you_are_vulnerable

[23] https://paragonie.com/blog/2015/05/using-encryption-and-authentication-correctly

Student Performance Analysis Using Educational Data Mining

P Ramya
M.Tech Student,
Gudlavalleru Engineering College,
Gudlavalleru, Krishna(Dt)
Vijayawada

M Mahesh Kumar
Asst Professor, Dept of IT
LakiReddy Balireddy College of Engineering,
Mylavaram, Krishna(Dt)
Vijayawada

***Abstract*— Software industry is hiring the students from the engineering colleges who are good in communication, programming, and also academically performing well. Most of the engineering institutions focused on the students performance on the above stated factors. The engineering students have to improve their academic performance, programming skills and also communication skills. To help such kind of students, we designed a project which can predict the students performance before the announcement of their results and before they attend their semester exams. By this the students can know their performance and can improve their skills by proper planning or by making changes in their plans. This can help the students improve in their academics, which eventually leads to a good performance in their end examinations. By this the suicide rates of students will also get reduced since the stress is reduced. This could help in our country development by providing good and efficient engineers to the country.**

We applying Naive Bayes classification algorithm and Weighted Naïve Bayesian algorithm on the student data set which is collected from LBRCE IT department, Mylavaram for building this model. Based on these results we can classify the weak students and take the remedial measures to improve their performance.

Keywords: Educational Data Mining, Classification, Prediction.

I. INTRODUCTION

The advent of information technology in various fields has lead the large volumes of data storage in various formats like records, files, documents, images, sound, videos, scientific data and many new data formats. The data collected from different applications require proper method of extracting knowledge from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data [1]. The main functions of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data [2]. Data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. Data mining techniques have been introduced into new fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc.

There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments [3]. Educational Data Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others. Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students‟ performance and so on.

The main aim of this project is to improvise the student performance in studies based on some important factors. Education is an essential element for the betterment and progress of a country. It enables the people of a country civilized and well mannered. Now-a-days developing new methods to discover knowledge from educational database in order to analyse student's trends and behaviours towards education. To analyse the data from different dimensions categorize it and to summarize the relationships. It motivated us to work on student dataset analysation.The data collection, categorization and classification is being performed manually. The main disadvantage of this process is delay in results, remedial measures are not taken properly due to late analysis of student performance. There will be delay in the results announcements which leads to the poor performance of the students in the next examination due to lack of planning in their preparation. When count of students increases, the analysis of performance of a student becomes difficult. To overcome this difficulty we now introduce you to educational data mining. When institutes store their students details in cloud, it will be difficult to analyse large data often called as big data. By applying data mining on the data stored, we can easily categories and analyse the results of a student in short time without any difficulties. Here, mainly concentrated on the students internal marks, ability to concentrate, attendance, awareness on course outcomes, tutorials, semester marks, content perception, assignments

II. DATA MINING DEFINITION AND TECHNIQUES

Data mining, also popularly known as Knowledge Discovery in Database refers to extracting or “mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The sequences of steps identified in extracting knowledge from data are shown in Figure 1.

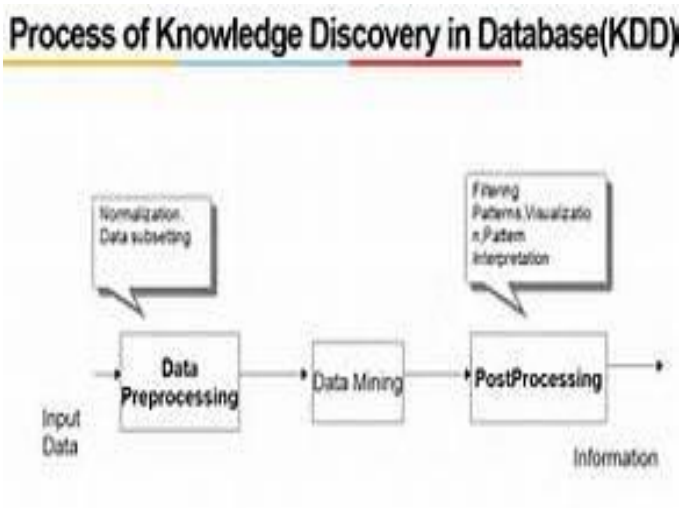


FIG 1:KDD PROCESS

Various algorithms and techniques like Classification , Clustering , Regression , Artificial Intelligence , Neural Networks , Association rules , Decision trees , Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. These techniques and methods in data mining need brief mention to have better understanding.

A. Classification

The Classification is the one of the most important technique used in data mining. It is a 2 step process 1.first build classification model. 2. Predict the class label, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach regularly employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

B. Clustering

Clustering can be defined as discovery of similar classes of objects.. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object
but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification.

C. Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

D. Association rule

Association and correlation is usually to find frequent item findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules confidence values less than one. However the number of Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if
any) value.

E. Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

F. Decision Trees

Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

G. Nearest Neighbor Method

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is greater than or equal to 1). Sometimes called the k-nearest neighbor technique.

III. Related Work

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes.

Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students as described by Alaa el-Halees [4]. Mining in educational environment is called Educational Data Mining.

Han and Kamber [3] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process.

Pandey and Pal [5] conducted study on the student performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

Hijazi and Naqvi [6] conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance" was framed. By means of simple linear regression analysis, it was found that the factors like mother's education and student's family income were highly correlated with the student academic performance.

Khan [7] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Galit [8] gave a case study that use students data to analyze their learning behavior to predict the results and to warn students at risk before their final exams.

Al-Radaideh, et al [9] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the NaïveBayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Pandey and Pal [10] conducted study on the student performance based by selecting 60 students from a degree college of Dr. R. M. L. Awadh University, Faizabad, India. By means of association rule they find the interestingness of student in opting class teaching language.

Ayesha, Mustafa, Sattar and Khan [11] describes the use of k-means clustering algorithm to predict student's learning activities. The information generated after the implementation of data mining technique may be helpful for instructor as well as for students.

Bray [12], in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Sri Lanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

Bhardwaj and Pal [13] conducted study on the student performance based by selecting 300 students from 5 different degree college conducting BCA (Bachelor of Computer Application) course of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayesian classification method on 17 attribute, it was found that the factors like students" grade in senior secondary exam, living location, medium of teaching, mother's qualification, students other habit, family annual income and student's family status were highly correlated with the student academic performance.

IV. Data Mining Process

In present day's educational system, a students" performance is determined by the internal assessment and end semester examination. The internal assessment is carried out by the teacher based upon students" performance in educational activities such as class test, seminar, assignments, general proficiency, attendance and lab work. The end semester examination is one that is scored by the student in semester examination. Each student has to get minimum marks to pass a semester in internal as well as end semester examination.

A. Data Preparations

The data set used in this study was obtained from LakiReddy Bali reddy College of Engineering ,Information Technology department, Mylavaram from session 2012 to 2016. Initially size of the data is 50. In this step data stored in different tables was joined in a single table after joining process errors were removed.

B. Data selection and transformation

In this step only those fields were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the database. All the predictor and response variables which were derived from the database are given in Table I for reference.

TABLE I. STUDENT RELATED VARIABLES

Variable	Description	Possible Values
IM	Internal Marks	{A>60% B>45 & <60% C>36 & <45% Fail<36%}
PSM	Previous Semester Marks	{A > 60% B >45 & <60% C >36 & <45% Fail < 36%}
Basics	Basics in the subject	{Poor , Average, Good}
ACIC	Ability to Concentrate in the Class	{Poor , Average, Good}
ASS	Assignment	{Yes, No}
CP	Content Perception	{Poor , Average, Good}
ATT	Attendance	{Poor , Average, Good}
Awareness on CO's	Course Outcomes Awareness	{Yes, No}
ESM	End Semester Marks	{First > 60% Second >45 & <60% Third >36 & <45% Fail < 36%}

The domain values for some of the variables were defined for the present investigation as follows:

Basics:

Helping students to study effectively. Easy to analyze the subject by knowing the basics and can easily remember the concept for longer time. Can generate new ideas. Allowing students to more clearly communicate ideas, thoughts and information. Helping students integrate new concepts with older concepts.

Ability to concentrate on the class:

Pay attention in the class is more important to gain more knowledge. Concentration in the class leads the students to understand the subject more easily. By paying attention in the class, students can do assignments & homework easily Can easily remember the topics being concentration in the classes. By taking notes in the class is helps to study easily. By concentration in the class students can take notes very effectively, which will help his/her further reference.

Attendance

The presence of student in a class can also improve his/her concentration in studies. Due to attendance marks ,the students attends the classes regularly .So, that they concentrate more in studies. Students can share knowledge with others. Can easily communicate with others.

Content perception:

By knowing about the content perception of a student, the teacher can help the student in understanding the subject further. We can assess whether the student listens or not by content perception.

Awareness on co's:

Before learning a subject one should have a clarity about what they are going to learn and why they are going to learn. When a student knows the course outcomes before starting the course, it will be easy for him/her to concentrate more on the subject. By having knowledge about course outcomes, the student gains interest to start that course and improve his knowledge.

Assignments:

By writing assignments, the students read the textbook, understand it and need to prepare notes for it. When a student frequently submits assignments, then the teacher can say that the student is regular and interested in learning by his/her own. By assignments, the students can learn subject by their own. Moreover, instead of reading subject, writing the subject improves the concentration of the student.

Internal marks:

The marks allotting to the students are divided as internal and external marks. The external marks are nothing but end exams (or) sem exams. By dividing the marks, makes it easier to assess the student performance more accurately. To assess our capability before end exams.

Semester Marks:

The semester marks of a student are helpful in analyzing performance of particular student. The semester marks are the marks that are obtained by a student in his/her end exam. The semester marks are converted in percentages and these percentages are considered during the campus placements as cut-off. The previous semester marks are considered to improve the students performance in their next semester. So that he can maintain percentage to get a good job. By considering the semester wise marks of a student, we can observe the change in the performance of that student.

Tutorials:

By conducting tutorials the staff (or) the teacher can maintain the record of a students performance. Observing the tutorials, the student can know where he should concentrate to score more marks.

C. Proposed System

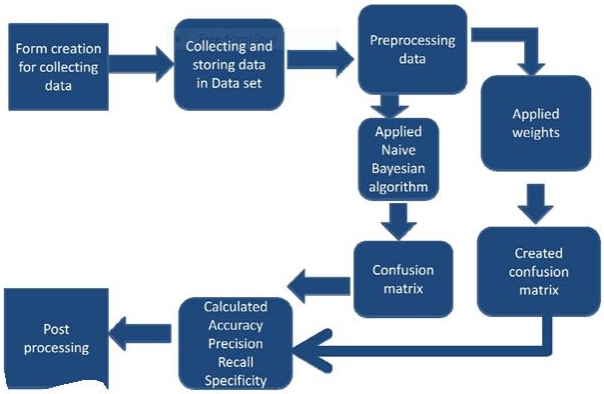


FIG2:BLOCK DIAGRAM

To justify the capabilities of data mining techniques in context of higher education by offering a data mining model for higher education system in the university we designed a model called "STUDENT PERFORMANCE ANALYSIS USING EDUCATIONAL DATA MINING". Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The main objective of this project is to use data mining methodologies to study student's performance in the courses. Data mining provides many tasks that could be used to study the student performance. In this project, the classification task is used to evaluate student's performance and as there are many approach that is used for data classification. Information's like Attendance, Class test, Seminar and Assignment marks were collected from the student's management system, to predict the performance at the end of the semester. This project reduces the time taken by the survey to collect the data, analyze the data and also reduces the errors in entering the data than that of the survey method. Software industry is hiring the students from the engineering colleges who are good in communication, programming, and also academically performing well. Most of the engineering institutions focused on the students' performance on the above stated factors. We are applying naive Bayes classification algorithm and weighted Naive Bayes algorithm on the student data set which is collected from LBRCE IT department, Mylavaram for building this model.

Modules include

1.Form Creation
2.Collection Of Trained Datasets
3.PreProcessing Datasets Collected
4.Applying Naive Bayesian Classifier
5.Applying Weighted Naive Bayesian Classifier
6.Calculating Confusion Matrix
7.Calculating Precision, Recall & Specificity
8.Comparision By Graphical Representation

V. Results and Discussion

The data set of 28 students used in this study was obtained from LakiReddy Bali eddy College of Engineering ,Dept of IT, Mylavaram from 2012 to 2016.

1.COLLECTION OF TRAINED DATASETS:

We created JSP files to store the trained dataset and test dataset into the database. Java Server Pages (JSP) is a technology for developing web pages that support dynamic content which helps developers insert java code in HTML pages by making use of special JSP tags. JSP is more powerful and easier to use. In the early days of the Web, the Common Gateway Interface (CGI) was the only tool for developing dynamic web content. However, CGI is not an efficient solution .JSP is the better solution for dynamic web content. Released in 1999 by Sun Microsystems, JSP is similar to PHP and ASP, but it uses java programming language.

SNo	ACO	Basics	ACOC	CP	IM	SM	ASS	TUT	ATT	PSM
1	No	Avg	Avg	Avg	A	Avg	No	Yes	A	Avg
2	Yes	Avg	Avg	Strong	C	Fail	No	No	C	Fail
3	No	Avg	Avg	Weak	C	Fail	No	No	B	Fail
4	No	Avg	Strong	Strong	B	Avg	No	No	B	Fail
5	No	Avg	Avg	Avg	B	Avg	Yes	No	C	Fail
6	No	Avg	Weak	Avg	C	Fail	No	No	B	Fail
7	No	Avg	Avg	Avg	A	Avg	Yes	No	A	Avg
8	No	Avg	Avg	Strong	A	Avg	Yes	Yes	A	Avg
9	No	Avg	Strong	Strong	C	Fail	No	No	C	Fail
10	No	Weak	Weak	Weak	C	Fail	No	No	C	Fail
11	No	Weak	Weak	Avg	B	Avg	No	Yes	B	Fail
12	No	Weak	Weak	Weak	D	Fail	No	No	C	Fail
13	No	Avg	Avg	Avg	B	Avg	Yes	Yes	B	Avg
14	Yes	Avg	Avg	Strong	C	Fail	No	No	C	Fail
15	No	Avg	Avg	Avg	A	Avg	No	Yes	A	Avg
16	Yes	Avg	Avg	Strong	C	Fail	No	No	C	Fail
17	No	Avg	Avg	Weak	C	Fail	No	No	B	Fail
18	No	Avg	Strong	Strong	B	Avg	No	No	B	Fail
19	No	Avg	Avg	Avg	B	Avg	Yes	No	C	Fail
20	No	Avg	Weak	Avg	C	Fail	No	No	B	Fail
21	No	Avg	Avg	Avg	A	Avg	Yes	No	A	Avg
22	No	Avg	Avg	Strong	A	Avg	Yes	Yes	A	Avg
23	No	Avg	Strong	Strong	C	Fail	No	No	C	Fail
24	No	Weak	Weak	Weak	C	Fail	No	No	C	Fail
25	No	Weak	Weak	Avg	B	Avg	No	Yes	B	Fail
26	No	Weak	Weak	Weak	D	Fail	No	No	C	Fail
27	No	Avg	Avg	Avg	B	Avg	Yes	Yes	B	Avg
28	Yes	Avg	Avg	Strong	C	Fail	No	No	C	Fail

TABLE 2: DATA SET OF STUDENTS

2. PRE-PROCESSING DATASETS COLLECTED:

Preprocessing is done in this module. Preprocessing techniques are data cleaning, data integration, data transformation, data reduction. In our project we are doing cleaning, transformation and reduction. In cleaning we are pruning incomplete values, inconsistent values and Null values. All these errors are pruned by using the java script. In transformation we are converting marks into grades to classify the end results.

3 APPLYING NAÏVE BAYESIAN CLASSIFIER:

In our project we are applying two algorithms on the student dataset to predict the student performance analysis. One is Naive Bayesian Algorithm and other is Weighted Naïve Bayesian Algorithm. Coming to Naïve Bayesian Algorithm, it is based on the Bayesian theorem. It is particularly suited when the dimensionality of the inputs is high. Parameter estimation for naive Bayes models uses the method of maximum likelihood. In spite over-simplified assumptions , it often performs better in many complex real-world situations.Advantage: Requires a small amount of training data to estimate the parameters. The Weighted Naïve Bayesian algorithm is also based on Bayesian theorem but some weights

are assigned to the attributes that plays a major role in the student end result prediction. When compared to Naïve Bayesian Algorithm , the weighted naive Bayesian algorithm gives more accurate results.

3.1 Naive Bayesian Classification Algorithm:

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes theorem. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large database. The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

Algorithm

Step-1: Let T be a training set of samples, each with their class labels. There are k classes,C1,C2…Ck. Each sample is represented by an n-dimensional vector,X={x1,x2,….xn}, measured values of n attributes, A1,A2,… An, respectively.

Step-2: Calculating prior probabilities

$p(Ci)=n(Ci)/m$ where $i=1,2,\ldots m$;

Step-3: Posterior probabilities
$p(Ci/X)=[p(X/Ci).p(Ci)]/p(X)$.

Step-4: Calculating
$p(X/Ci)=n\pi k=1p(Xk/Ci)$.

Step-5: In order to predict the class label of X,p(X/Ci)p(Ci) is evaluated for each class Ci.

$p(X/Ci).p(Ci)>p(X/Cj).p(Cj)$ for $1\leq j\leq m, j\neq i$.

3.2 Weighted Naïve Bayesian Classification Algorithm : 74

The weighted Naive Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. In this we are applying weights for each and every attribute.

Algorithm

Step-1: Let T be a training set of samples, each with their class labels. There are k classes,C1,C2…Ck. Each sample is represented by an n-dimensional vector,X={x1,x2,….xn}, measured values of n attributes, A1,A2,… An, respectively.

Step-2: Calculating prior probabilities

$p(Ci)=n(Ci)/m$ where $i=1,2,\ldots m$;

Step-3: We add weights to Xn if their value is highest among the values.
[P(X/Ci).P(Ci)] + Value.

Step-4: Else we do calculation
$p(X/Ci)=n\pi k=1p(Xk/Ci)$.

Step-5: In order to predict the class label of X,p(X/Ci)p(Ci) is evaluated for each class Ci.
$p(X/Ci).p(Ci)>p(X/Cj).p(Cj)$ for $1\leq j\leq m, j\neq i$.

Weights Included

The below table contains Boolean value attribute weights from scale 0-1. These weights are added in weighted naïve Bayesian algorithm, so that to get more accurate results than that of naïve Bayesian classifier. The Boolean valued attributes are nothing but having binary values like yes or no, true or false.

Boolean value Attribute weights from 0-1 scale						
S.No	Awareness of CO's		Assignments		Tutorials	
	Yes	No	Yes	No	Yes	No
Professor 1	0.18	0.0	0.22	0.0	0.22	0.0
Professor 2	0.20	0.0	0.18	0.0	0.18	0.0
Professor 3	0.22	0.0	0.20	0.0	0.20	0.0
Average	0.20	0.0	0.20	0.0	0.20	0.0

TABLE 3: BOOLEAN VALUE ATTRIBUTE WEIGHTS

Multi value Attribute weights from 0-1 scale									
S.No	Basics			Ability to Concentrate in the Class			Content Perception		
	S	Avg	W	S	Avg	W	S	Avg	W
Professor 1	0.50	0.18	0.0	0.60	0.32	0.0	0.75	0.50	0.0
Professor 2	0.45	0.22	0.0	0.55	0.28	0.0	0.85	0.48	0.0
Professor 3	0.65	0.20	0.0	0.65	0.30	0.0	0.80	0.52	0.0
Average	0.50	0.20	0.0	0.60	0.30	0.0	0.80	0.50	0.0

S:Strong Avg:Average W:Weak

TABLE 4: MULTI VALUE ATTRIBUTE WEIGHTS

Multi Value Attribute Weights				
S.No	A	B	C	D
Professor 1	0.90	0.72	0.50	0.0
Professor 2	0.88	0.68	0.8	0.0
Professor 3	0.92	0.70	0.52	0.0
Average	0.90	0.70	0.50	0.0

TABLE 5: MULTI VALUE ATTRIBUTE WEIGHTS

NOTE:WE DON'T TAKE ANY WEIGHTS FOR ATTENDANCE ATTRIBUTE

CALCULATING CONFUSION MATRIX

		PREDICTED	
		Negative	Positive
Actual	NEGATIVE	A	B
	POSITIVE	C	D

TABLE 6 CONFUSION MATRIX

The entries in the confusion matrix have the following meaning in the context of our study:

a. is the number of correct predictions that an instance is negative,
b. is the number of incorrect predictions that an instance is positive,
c. is the number of incorrect of predictions that an instance negative, and
d. is the number of correct predictions that an instance is positive

CALCULATING RECALL,PRECISION & SPECIFICITY

In this module the accuracy, precision, recall and specificity are calculated from the confusion matrix. By considering the above table accuracy, precision, recall and specificity are defined below.

Several standard terms have been defined for the 2 class matrix:
The accuracy (AC) is the proportion of the total number of predictions that were correct.
It is determined using the equation

AC= (a+d)/(a+b+c+d).

The recall or true positive rate (TP) is the proportion of positive cases that were correctly
It is determined using the equation

TP=d/(c+d),

The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive as calculated using the formula

FP= b/(a+b).

The true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly as calculated using the equation.

TN= a/(a+b).

The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative as calculated using the equation.

FN= c/(c+d).

Finally, precision (P) is the proportion of the predicted positive cases that were correct, as calculated using the equation.

P=d/ (b+d).

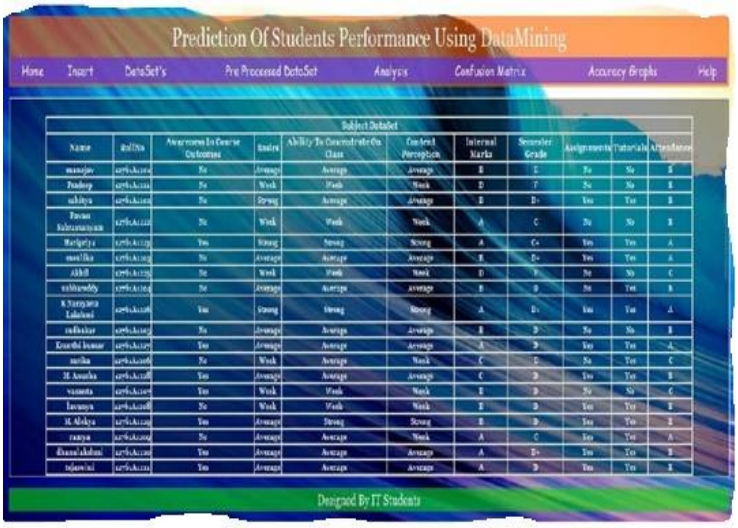


Fig 3:Data set of 28 Students

Fig 4:Preprocessed data set of collected Students data set

Fig 5:Comparison of Bayesian and Weighted Bayesian Classifier

CONCLUSION

In this paper, the classification task is used on student database to predict the students division on the basis of previous database. As there are many approaches that are used for data classification, the Naïve Bayesian Classifier and Weighted Naïve Bayesian Classifier are used here. Information's like Attendance, Class test, Seminar and Assignment marks were collected from the student's previous database, to predict the performance at the end of the semester.

This study will help to the students and the teachers to improve the division of the student. This study will also work to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination. This can help the students improve in their academics, which eventually leads to a good performance in their end examinations. By this the suicide rates of students will also get reduced since the stress is reduced. This could help in our country development by providing good and efficient engineers to the country.

REFERENCES

[1] Heikki, Mannila, Data mining: machine learning, statistics, and databases, IEEE, 1996.

[2] U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6,1996.

[3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study", 2009..

[4] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[5] Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6,1996.

J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[6] Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A

[7] Case Study", 2009..Morgan Kaufmann, 2000.

[8] Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study", 2009..

[9] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[10]

Case Study", 2009..

[11] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[12] Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6,1996.

[13] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[14] Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study", 2009..

[15] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[16] S. T. Hijazi, and R. S. M. M. Naqvi, "Factors affecting student's performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[7] Z. N. Khan, "Scholastic achievement of higher secondary students in science stream", Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005..

[8] Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education 2007.

[9] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, "Mining student data using decision trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[10] U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[11] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data mining model for higher education system", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.

[12] M. Bray, The shadow education system: private tutoring and its implications for planners, (2nd ed.), UNESCO, PARIS, France, 2007.

[13] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[14] J. R. Quinlan, "Introduction of decision tree: Machine learn", 1: pp. 86-106, 1986.

[15] Vashishta, S. (2011). Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm. *IJACSA - International Journal of Advanced Computer Science and Applications*, *2*(4), 77-80.

[16] Kumar, V. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. *IJACSA - International Journal of Advanced Computer Science and Applications*, *2*(3), 80-84. Retrieved from http://ijacsa.thesai.org.

Detecting and Preventing CSRF Attack on Web Application

N.Vidya Rani , Dr. G. Ramakoteswara rao
Department of Information technology,
VR Siddhartha Engineering College,
Vijayawada, Andhra Pradesh, India

***Abstract*— Cross-Site Request Forgery attacks occur when malicious web site causes the user's browser to perform any action on a trusted site. These attacks have been called the "sleeping giant" of web-based vulnerabilities, as many websites do not protect them and because they have been largely ignored by the web development and security communities. We present one serious CSRF vulnerabilities we have discovered an important site, including what we believe is the first published attack intervention of a financial institution. We recommend changes to the server that is able to completely protect a site from CSRF attacks. The characteristics of a server solution must have also described. In addition, we have implemented a browser plug-in client side, which can protect against certain types of CSRF attacks, even if the site does not take steps to protect it. We hope to attract the attention of CSRF attacks while those responsible developers of web tools to protect users against these attacks.**

Keywords—cross-site request forgery,prevention,database,web.

I. INTRODUCTION

CSRF is an attack which forces an end user to execute unwanted actions on a web application that are currently authenticated actions. CSRF attacks specifically target status change requests, no data theft, because the attacker has no way to see the response to the request forged. With a little help of social engineering, an attacker can trick users of a web application in the implementation of actions of the attacker's choice. If the victim is a regular user, the successful CSRF attack can force the user to request a change of status as the transfer of funds to change their email address, and so on. If the victim is an administrative account, CSRF could jeopardize the entire web application.

CSRF is an attack that tricks the victim to file a malicious application The identity and privileges of the victim to perform an unwanted on behalf of the victim role is inherited. For most sites, the browser automatically includes requesting credentials associated with the site, such as user session cookies, IP-address, domain credentials Windows, and so on. Therefore, if the user is now authenticated to the site, the site will have no way of distinguishing between forged request sent by the victim and a legitimate request sent by the victim. CSRF attacks target function, which causes a change in state on the server, for example, change your email address or password victim, or buy something. Forcing the victim to extract data does not benefit the attacker because the attacker does not receive a response, the victim does. Thus, CSRF attacks change requests are sent to the state. Sometimes you can save CSRF attack on the most vulnerable site. These vulnerabilities are called "stored CSRF flaws." This can be achieved simply store an IMG or IFRAME tag in a field that accepts HTML, or by a cross-site scripting attack more complex. If the attack can store a CSRF attack on the site, the severity of the attack is amplified. In particular, the probability increases because the victim is more likely to view the page containing the attack some random page on the Internet. The probability is also increased because the victim is sure to be authenticated on the site already. CSRF attacks are also known by several names, including XSRF, "Surf Mar", section management, including reference sites Forgery, and the hostile link. Microsoft refers to this attack as an attack by a click on your threat modeling process and many places in its online documentation.

II. RELATED WORKS

The impact of CSRF attacks got it thanks largely to the work of Chris Shiflett [7] OmniTI and Jeremiah Grossman [2] WhiteHat security. Burns [2] and Schreiber [6] providing comprehensive introductions to attack CSRF, but do not describe the vulnerability works. Jones and winter [3] describe RequestRodeo, client-side protection from CSRF attacks using HTTP-Proxy. This approach has some limitations and descriptions browser, similar to ours, as a possible future of the plugin. Expanded job in [4] by implementing a LAN is limited to prevent CSRF attacks on local resources.

There are server's remedies that are similar to our recommendations, but no standard requirements caused unnecessary problems. As mentioned, Jones and winter [3] and Schreiber [6], require that the server status, while Shiflett [7] broken navigation tabs. Jovanovich [5] created a method of adapting legacy applications with CSRF protection by adding a proxy between the web server and web applications. These security measures require that all data in the buffer and link the modified application. The challenges also require some programs will be rewritten. This decision shall enter into force when the native application cannot be rewritten, but not as effective as adding protection from CSRF directly to the application. This solution is intended for administrators who want to protect their applications on servers CSRF attacks, while our solution is designed for developers of web applications that want to add a frame, and CSRF protection directly to their programs.

III. Overview of CSRF

Figures 1, 2 and 3 show how CSRF attacks generally work. Then CSRF attacks are described in more detail using a specific example.

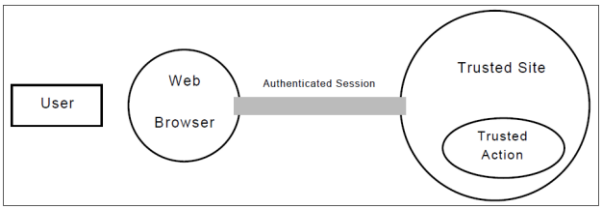


Figure 1: The web browser has established a session authenticated with the trusted site. Confidence action should only be performed when the web browser makes the request through the authenticated session.

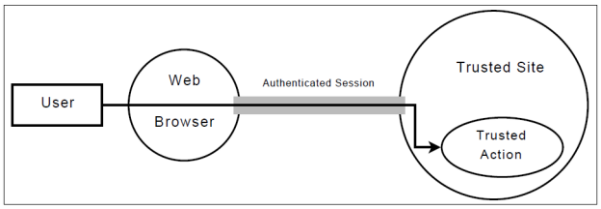


Figure 2: A valid application. The Web browser attempts to perform an action of confidence. The trusted site confirms that the Web browser is authenticated and allows the action to perform.

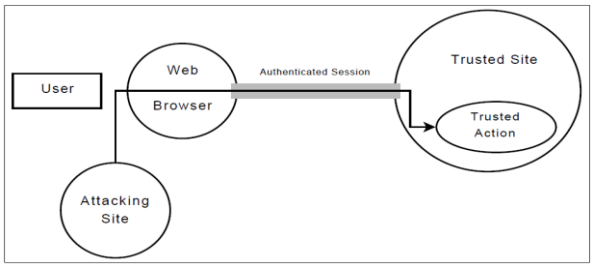


Figure 3: A CSRF attack. The site Attacking causes the browser to send a request to the trusted site. The trusted site sees a valid application, authenticated from the web browser and performs reliable action. CSRF attacks are possible because websites web browser, not the user is authenticated.

A. An Example

Consider the hypothetical example of a site vulnerable to attack CSRF. This site is a website based email web that allows users to send and receive e-mail. The site uses Digest authentication to authenticate users. On the page, http://example.com/compose.htm form contains HTML, which allows the user to enter an email address, the subject and the message, and the button that says "Send Message."

```
< form
action = "http://localhost:8088/send_email.htm"
method = "GET" >
Recipient's E − Mail address: < input
type = "text" name = "to" >
Subject: < input type = "text" name = "subject" >
Message: < textarea name = "msg" ></textarea >
< input type = "submit" value = "Send Email" >
</form >
```

When a user clicks example.com "Send email", the data entered will be sent to http://example.com/send_email.htm as a GET request. From a GET request, simply add the form data to the URL, the user will be sent to the following URL? :

```
http://localhost:8088/send_email.htm?to = bob%40example.com&subject = hello&msg
= What%27s + the + status + of + that + proposal%3F 3
```

Page send email.htm would take the data it receives and sends an email to the recipient by the user. Note that send email.htm simply takes data and performs an action with that data. No matter where the request originated, only that the request was made. This means that if the users manually entered into the above address into your browser, example.com still have to send an email. For example, if the user typed the following three URLs in your browser, send email.htm send three emails:

```
http://localhost:8088/send_email.htm?to =
bob%40example.com&subject = hi + Bob&msg = test}

http://localhost:8088/send_email.htm?to =
alice%40example.com&subject = hi + Alice&msg =
test}

http://localhost:8088/send_email.htm?to =
carol%40example.com&subject = hi + Carol&msg =
test}
```

CSRF attack here, as Send email.htm receives the data it receives and sends e-mail. It does not check that the data came from the way compose.htm. Thus, if an attacker can cause the user to send a request to send email.htm, which will send a page the user name example.com email containing information about elections attacker and the attacker made a successful attack will CSRF.

To take advantage of this vulnerability, an attacker must force the user's browser to send a request to send email.htm to perform some nefarious actions. Specifically, the attacker has to create a request through the site your site for example.com. Unfortunately, HTML provides many ways to make such requests. <IMG>, for example, force the browser to download any URI is set as Src attribute, even if the URI is not an image. An attacker could create a page with the following code:

```
< img src = "http://example.com/send_email.htm?to = mallory%40example.com&subject = Hi&msg
= My + email + address + has + been + stolen" >
```

When the user visits the page, a request to send email.htm be sent, which in turn send an email to Mallory by the user.

CSRF attacks are successful when an attacker can cause the user's browser to perform an unwanted action elsewhere. For this action to be successful, the user should be able to do this. CSRF attacks are typically so powerful as a user, i.e. any action that the user can also be performed by an attacker using a CSRF attack. Consequently, the increased power of a site gives a user, the more severe the possible CSRF attacks. CSRF attacks can succeed against almost all sites using implicit and explicit authentication does not protect against CSRF attacks. The same origin policy was designed to prevent an attacker to access data on a third party site. This policy does not prevent Applications sent, it only avoids an attack from analysis the data returned by the third-party server. Since CSRF attacks are the result of requests sent, the same origin policy does not protect against CSRF attacks.

B. Authentication and CSRF

CSRF attacks often exploit the authentication mechanisms targeted sites. The root of the problem is that the Web authentication usually ensures a site that a request came from the browser of a specific user; but it does not guarantee that the user actually requested or authorized the request.

For example, suppose Alice visits a target site T. T Alice gives the browser a cookie contains a session identifier sid pseudorandom, to keep track of your session. Alice is asked to log on to the site, and input your username and valid password, the site records the fact that Alice is recorded in the sid session. When Alice sends a request to T, your browser automatically sends the session cookie containing sid. T then uses your registry to identify the session as coming from Alice.

Now suppose Alice visits a malicious site content supplied by M. M contains Javascript code or an image tag that makes Alice's browser to send an HTTP request to T. Because the application is going to T, the browser Alice "kindly" adds the SID session cookie to the request. Seeing the request, T follows from the presence of the cookie that the request came from Alice, so T performs the requested process on Alice's account. This is a positive CSRF attack.

Most of the other Web authentication mechanisms suffer from the same problem. For example, the mechanism BasicAuth HTTP [22] would Alice tell your browser your username and password to the site of T, then the browser would be "kindly" to unite the username and password for future requests sent to T . Alternatively, T may use the SSL client certificates -SIDE, but the same problem would result because the browser would be "kindly" to use the certificate to carry out requests to the site of T. Similarly, if T authenticates Alice by your IP address, CSRF attacks would be possible.

In general, every time you pass Digest authentication, so a site request is being sent to the browser, which is coming there is a danger of CSRF attacks. In principle, this danger could be abolished by requiring the user to take an obvious, un-spoof able action for each request sent to a site, but in practice this would cause major usability problems. Standard authentication mechanisms most widely used and do not prevent CSRF attacks, so that a practical solution must be sought elsewhere.

C. CSRF Attack Vectors

For an attack to be successful a user must log on to the destination site and to visit the attacker's site or a site on which the attacker has partial control.

If the server contains vulnerabilities CSRF and also accepts GET requests, CSRF attacks are possible without the use of JavaScript. If the server only accepts POST requests, JavaScript is required to automatically send a POST request from the attacker's site to the target site.

D. CSRF vs. XSS

Recently, much attention has been paid to Cross-Site Scripting (XSS) [20] vulnerabilities. XSS attack occurs when an attacker introducing malicious code into a site to be assigned to other users. For example, a site may allow users to leave comments. These comments are sent by a user, stored in a database and all future users of the site are displayed. If an attacker is able to enter malicious JavaScript as part of a comment, the JavaScript code would be embedded in any page containing the comment. When a user visits the site, JavaScript attacker would run with all the privileges of the target site. Malicious JavaScript embedded in a target site would be able to send and receive applications from any page and access cookies set by this site. XSS protection required to filter sites carefully any user input to make sure that no malicious code is injected.

CSRF and XSS attacks differ in that the XSS attacks require JavaScript, while CSRF attacks do not. XSS attacks require sites accept malicious code, while CSRF attacks with malicious code is on third party sites. Filtering user input will prevent the execution of malicious code on a site, but it does not hurt to launch malicious code on other sites. Since the malicious code can run on third-party sites, protection against XSS does not protect a site from CSRF attacks. If a site is vulnerable to XSS, then it is vulnerable to CSRF attacks. If a site is completely protected from XSS attacks are more likely it remains vulnerable to CSRF attacks.

IV. CSRF Vulnerabilities and Prevention

Here we describe a vulnerability we found. These attacks were found by the survey list of popular web services. The fact that many websites are vulnerable to CSRF attacks, while third party reveals the problem shows that many site administrators ignorant of the risks and the existence of CSRF vulnerabilities.

We found CSRF vulnerabilities in web applications, allowing banking attacker access to more accounts in the name of the user and transfer funds from an account user on account of the attacker. The banking application using SSL prevents this attack. We believe that this is the first published CSRF attack involving financial institution. As banking application was clearly not protect against CSRF attacks, transferring funds from the user's account was as simple as mimicking the steps a user must take when transferring funds. These steps consist of the following:

1. The attacker creates a user account to access banking site.

a. The attacker causes the user's browser to visit banking application "Create New User Account" page:

A GET request to http://localhost:8088/CSRF/registerUser.jsp?acno=7&name=user7&address=user7&city=user7&email=user7&mobile=998877665544&pass=user7&cpass=user7

2. The attacker starting transactions with user account.

a. The attacker depositing amount:

A GET request to deposit amount

http://localhost:8088/CSRF/Deposit1.jsp?amount=10000&userid=hari&submit=Deposit

b. The attacker transferring funds to other account:

A GET request to transfer funds to other account http://localhost:8088/CSRF/Transfer1.jsp?amount=1000&userid=hari&to=00001&submit=Transfer

To exploit this attack, an attacker would create a page that made the above POST requests in succession using JavaScript. This would be invisible to the user. This attack assumes the user has not added an additional payee to his banking application checking account. The attack could likely have been modified to work without this restriction.

V. Conclusion

CSRF attacks are relatively easy to diagnose, exploit and correct. The most plausible explanation for the prevalence of these attacks explanation is that web developers are unaware of the problem or think defenses against the best-known cross-site scripting attacks also protect against CSRF attacks. We hope that the attacks we have presented show the danger of CSRF attacks and web developers helps give these attacks the attention they deserve. The root cause of CSRF and similar vulnerabilities probably lies in the complexity of Web protocols today, and the gradual evolution of the Web from a facility for submitting data to a platform for interactive services. The more opportunities added to the client browser and the more complex sites include interactive services and client-server programming and related CSRF attacks become more frequent if the protection will not be accepted. As the complexity of web technologies continue to grow, we can expect more attacks new categories emerge.

REFERENCES

[1] J. Burns. Cross Site Reference Forgery: An introduction to a common web application weakness. http://www.isecpartners.com/documents/XSRF_Paper.pdf, 2005.

[2] J. Grossman. CSRF, the sleeping giant. http://jeremiahgrossman.blogspot.com/2006/09/csrf-sleeping-giant.html,Sep 2006.

[3] M. Johns and J. Winter. RequestRodeo: Client Side Protection against Session Riding. In F. Piessens, editor, Proceedings of the OWASP Europe 2006 Conference, refereed papers track, May 2006.

[4] M. Johns and J. Winter. Protecting the Intranet against "JavaScript Malware" and Related Attacks. In DIMVA, 2007.

[5] N. Jovanovic, E. Kirda, and C. Kruegel. Preventing Cross Site Request Forgery Attacks. Securecomm and Workshops, 2006, Sept 2006.

[6] T. Schreiber. Session Riding: A Widespread Vulnerability in Today's Web Applications. http://www.securenet.de/papers/Session_Riding.pdf, 2004.

[7] C. Shiflett. Security Corner: Cross-Site Request Forgeries. http://shiflett.org/articles/cross-site-requestforgeries,2004.

REVIEW ON BASTION HOSTS

G.Vijayababu[1], D.Haritha[2], R.Satya Prasad[3]
[1] Research Scholar, Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, INDIA
[2] Professor, S.R.K. Institute of Technology, Vijayawada, Andhra Pradesh, INDIA
[3] Associate Professor, Acharya Nagarjuna University, Guntur, Andhra Pradesh, INDIA
g.vijayababu777@gmail.com, harithadasari@rediffmail.com, profrsp@gmail.com

Abstract

Bastion hosts play a significant role in providing secure information flow between the private and public networks and thus secures the internal network from the external intruders. Bastion hosts sit on the network perimeter and can play several roles of Bastion hosts like router, DNS, FTP, SMTP, News, and/or Web servers. The different configurations and locations that the bastion hosts can be hosted make difference in their performance in different situations. This paper discusses the various bastion host architectures and building of bastion host in detail. The significance of the bastion host hardening is described.

Keywords: Bastion Host ,DMZ, Firewall, Bation host hardening

1.INTRODUCTION

Bastion Hosts are designed for secure information flow between public network and private network. Bastion hosts sit on the network perimeter. Bastion host is a server and it is meant to provide access to a private network from an external network, like Internet. The system is on the public side of the demilitarized zone (DMZ). The hardening of Bastion hosts resists attacks from external sources thus protecting the internal network. Hardening involves securing the machine, configuring the required services, installing the necessary patches, controlling the services and protocols, locking the user accounts via defining and modifying the Access Control Lists (ACLs),disabling all unnecessary TCP and UDP ports and running the security audit to establish a baseline. Frequently all these functionalities are critical to the network security system. Typically, the bastion host serves acts as a platform for an application level gateway (proxy) or circuit level gateway.

.

The application level gateway (proxy) is a firewall and it allows users to run specific service like FTP, TELNET, HTTP etc. or specific connection by implementing authentication, filtering and logging. Each specific service has it's own specific proxy. For example, if only HTTP connection is allowed to the Internet for internal network users, then HTTP proxy must be allowed, no other proxy is allowed. Users who need to go to Internet create a virtual circuit with the proxy server and the proxy server sends the request to connect to a specific site. Proxy server protects or hides the internal network by sending the request with its own IP. Only the IP of the proxy server is visible to the external world. After receiving the response from the Internet it sends it back to its intended internal user via the virtual circuit. The proxy is aware of the type of data it handles and can give protection to it.

The circuit-level gateway can be a stand-alone system or an application-level gateway with special functionality for certain applications. As with an application gateway, a circuit-level gateway does not permit an end-to-end TCP connection. Instead it makes two TCP connections, one between TCP user on the inner host and itself and one between itself and a TCP user on an outside host. Once the two connections are established, the gateway typically relays TCP segments between those two connections without examining the contents. The Gateway secures by accurately determining which connections are to be allowed.

Circuit-level gateway is typically used when the system administrator trusts the internal users. The gateway can be configured to support application- level or proxy service on inbound connections and circuit-level functions for outbound connections. In this configuration, the gateway can incur the overhead of examining incoming application data for forbidden functions but does not incur that overhead on outgoing data.The bastion host hardware platform takes support of a secure version of operating system and makes it a hardened system.

Only the essential services that are considered by the network administrator are installed on the bastion host which include proxy applications for DNS,FTP, HTTP, and SMTP. Prior to allowing access to the proxy services by the user, the bastion host may require additional authentication. In addition, each proxy service may require its own authentication before granting user access. The configuration of each proxy is done in such a way that it supports only a subset of the standard application's command set. The configuration of each proxy is done to allow access only to specific host systems. Detailed

audit information is maintained by each proxy by logging all traffic, each connection, and the duration of each connection. It is an essential tool for identifying and terminating intruder attacks.

As each proxy module is a tiny software package which is specifically designed for the security of network, it is easier to check for security flaws. Each proxy is independent of other proxies on the bastion host. Any proxy module can be easily uninstalled without affecting the other proxy modules in case of any problem, or if a future vulnerability is discovered. Similarly a new proxy service can be easily installed on the bastion host.

A proxy do not access disk except to read its initial configuration file. By making the portions of the file system containing executable code as read only, makes difficult for an intruder to install Trojan horse sniffers or other dangerous files on the bastion host. Each proxy runs as a non privileged user in a private and secured directory on the bastion host. Bastion hosts only permit logins at the system console and does not permit network logins.

In unlabelled class of bastion host, the configuration should be strictly controlled and restricted to strictly necessary features, there by many of the inherent vulnerabilities of the base operating system will be avoided, though some will still be present. If a proxy can be hijacked, a hacker can use standard hacking techniques to gain entry to the bastion host configuration and the internal network that the bastion host is protecting. Gauntlet from Trusted Information Systems is such a bastion host. Bastion hosts of this type are being evaluated to ITSEC E3 assurance, though if the underlying operating system is only evaluated to ITSEC E2, the exact amount of assurance gained is uncertain.

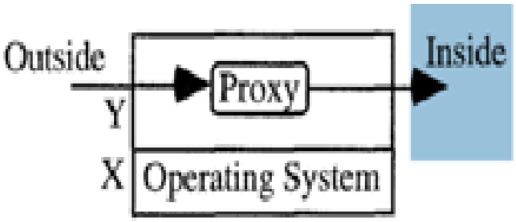


Fig 1.1 Bastion host on Labelled OS

Bastion hosts mounted on labelled operating systems are aimed at both the commercial and government markets As shown in figure 1.1 the proxies execute with a process label of Y, where as the operating system and bastion host configuration files are labelled X. These labels are arbitrary labels, and are related by the fact that Y strictly dominates **X.** Because of this relationship, the proxies can read the configuration files, but not write to them. Like the firewalls based on unlabelled operating systems, some of the proxies that are supplied with these firewalls are evaluated, although not all of them are. If a fault is found in an evaluated proxy, then any fix technically invalidates the evaluation and the proxy must be reevaluated.[2]

A successful exploitation of the fault is also likely to result in an unauthorized entry into the inside network, since it is only the functionality of the proxy that offers any defense. Consistency, completeness, and compactness need to be considered while designing the rules of the Firewall [7].

Section 2 describes different standard Bastion host configurations. Section 3 discusses some common variations, and their benefits and drawbacks. In Section 4 Bastion host generic architecture is given. Bastion host hardening procedure is discussed in Section 5 with conclusions in Section 6.

2. BASTION HOST FIREWALL CONFIGURATIONS

There are two types of screened host-one is single homed bastion host and the other one is dual homed bastion host. In case of single homed bastion host the firewall system consists of a packet filtering router and a bastion host. A bastion host is basically a single computer with high security configuration, which has the following characteristics:

- Traffic from the Internet can only reach the bastion host; they cannot reach the internal network.
- Traffic having the IP address of the bastion host can only go to the Internet. No traffic from the internal network can go to the Internet.

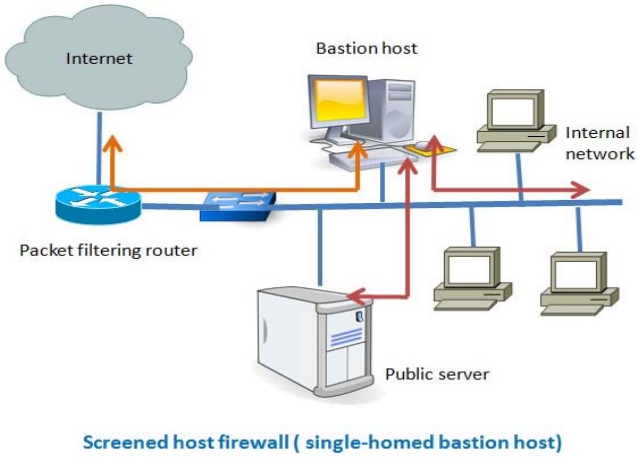


Fig 2.1 Screened host firewall (Single Homed Bastion Host)

This type of configuration can have a web server placed in between the router and the bastion host in order to allow the public to access the server from the Internet. The main problem with the single homed bastion host is that if the packet filter route gets compromised then the entire network will be compromised. To eliminate this drawback the dual homed bastion host firewall system can be used.Where a bastion host has two network cards- one is used for internal connection and the second one is used for connection with the

router. In this case, even if, the router got compromised, the internal network will remain unaffected since it is in the separate network zone.

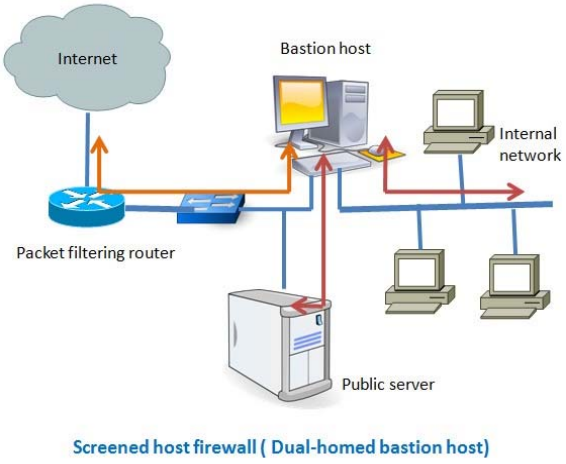


Fig 2.2 Screened Host Firewall (Dual Homed Bastion Host)

Screened subnet firewalls

This is one of the most secured firewall configurations. In this configuration, two packet filtering routers are used and the bastion host is positioned in between the two routers. In a typical case, both the Internet and the internal users have access to the screened subnet, but the traffic flow between the two subnets (one is from bastion host to the internal network and the other is the sub-network between the two routers) is blocked.

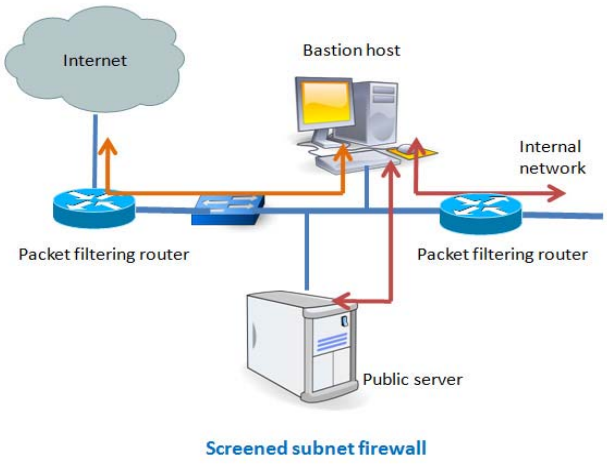


Fig 2.3 Screened subnet firewall

Packet filtering Router functions as a firewall by examining every packet passing through the network. Based on access control list, the router either forward or drop packets. Normally, the IP address of the source and destination, port number and type of traffic are taken into account when the router processes each data packet. Since a router cannot check packet in the application layer, this type of firewall cannot defend attacks that use application layers vulnerabilities.

3. VARIATIONS ON FIREWALL ARCHITECTURES

The most common firewall architectures are shown. However, there is a lot of variation in architectures. There is a good deal of flexibility regarding how to configure and combine firewall components to best suit the intended hardware, budget, and security policy. These variations are discussed here[1].

3.1 Usage of multiple bastion hosts

It does make sense to use multiple bastion hosts in firewall configuration, as shown in Figure 3.1 . Reasons for doing this include performance, redundancy, and the need to separate data or servers.

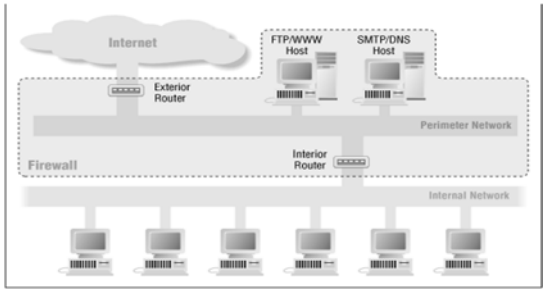


Figure 3.1: Architecture using two bastion hosts

If it is decided to have one bastion host to handle the services that are important to users (such as SMTP servers, proxy servers, and so on), while another host handles the services that are provided to the Internet, but which the users don't care about (for example, an anonymous FTP server). In this way, performance for the users won't be degraded by the activities of outside users.

Multiple bastion hosts can be employed with the same services for improving performance, but it's difficult to do load balancing. Most services need to be configured for particular servers, so creating multiple hosts for individual services works best if the prediction of usage is made in advance.

If the firewall configuration includes multiple bastion hosts, they might be configured for redundancy, so that if one fails, the services can be provided by another, but it should be carefully noted that only some services support this approach. For example, multiple bastion hosts might be configured and designated as DNS servers for particular domain (via DNS NS [Name Server] records, which specify the name servers for a domain), or as SMTP servers (via DNS MX [Mail Exchange] records, which specify what servers will accept mail for a given host or domain), or both. Then, if one of the bastion hosts is unavailable or overloaded, the DNS and SMTP activity will use the other as a fallback system.

Multiple bastion hosts might be used to keep the data sets of services from interfering with each other. Security is another factor for this separation in addition to the performance. For example, it might be decided to provide one HTTP server for use by a particular group of customers over the Internet and another for use by all others. By providing two servers, different data can be offered to customers with possibly better performance, by a powerful machine or a machine with less load.

HTTP server and anonymous FTP server can be run on separate machines, to eliminate the possibility that one server could be used to compromise the other.

3.2 Merging the Interior Router and the Exterior Router

The interior and exterior routers can be merged into a single router, provided there is a router sufficiently capable and flexible. In general, a router that allows specifying both inbound and outbound filters on each interface is needed.

If the interior and exterior routers are merged, as shown in Figure 3.2 , still a perimeter net (on one interface of the router) and a connection to internal net (on another interface of the router)might be there. Some traffic handled by proxies flow between internal net and perimeter net, some traffic flow between perimeter net and internet and the traffic that is permitted by the packet filtering rules set for the router flow directly between internal net and internet.

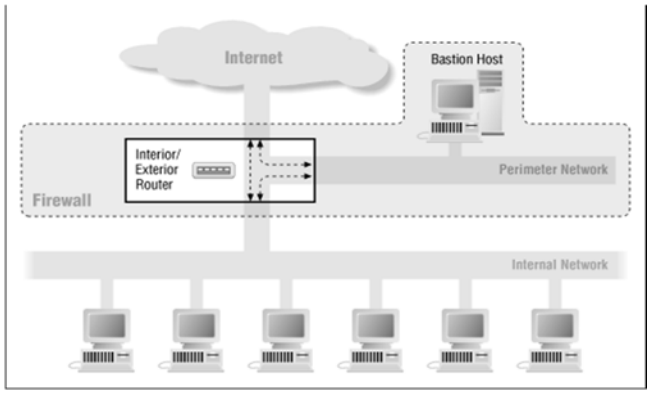


Figure 3.2: Architecture using a merged interior and exterior router

This architecture, like the screened host architecture, makes the site vulnerable to the compromise of a single router. In general, routers are easier to protect than hosts, but they are not impenetrable.

3.3 Merging the Bastion Host and the Exterior Router

There might be cases in which a single dual-homed machine is used as both bastion host and exterior router. Here's an example: suppose there is only a dial-up SLIP or PPP connection to the Internet. In this case, something like the Morning Star PPP package can be run on the bastion host, and allowed to act as both bastion host and exterior router. It is equivalent to have the three-machine configurations bastion host, interior router and exterior router.

Using a dual-homed host to route traffic won't give the performance or the flexibility of a dedicated router. Depending on the operating system and software that is being used, there may or may not be the ability to do packet filtering. Many interface software packages like Morning Star PPP package have quite good packet filtering capabilities. However, the exterior routers doesn't have to do much packet filtering, using an interface package with average packet filtering capabilities is not a big problem.

Merging the bastion host with the exterior router as shown in Figure 3.3 does not cause significant new vulnerabilities unlike the merging of internal and external routers. In this architecture, the bastion host is more exposed to the Internet, protected only by its own filtering capabilities and thus extra care is needed to protect it.

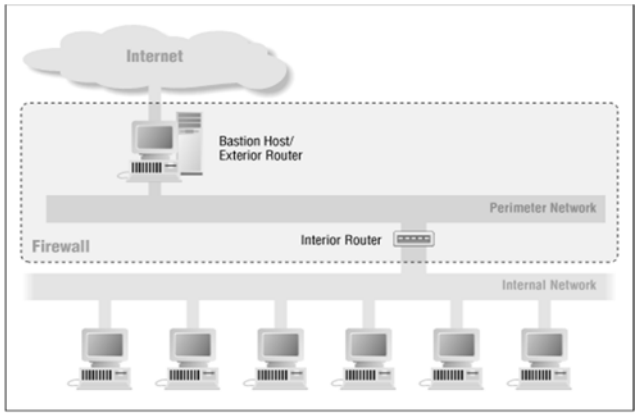


Figure 3.3: Architecture using a merged bastion host and exterior router

3.4 Dangers of Merging the Bastion Host and the Interior Router

Unlike merging the bastion host and the exterior router, it's not a good idea to merge the bastion host and the interior router, as shown in Figure 3.4. Doing so compromises the overall security. The bastion host and the exterior router perform distinct protective tasks but the interior router functions in part sas a backup to the two of them. If the bastion host and the interior router are merged, then the firewall configuration is changed in a fundamental way. With a separate bastion host and interior router, screened subnet firewall architecture is available. The perimeter net for the bastion host doesn't carry any internal traffic. Even if the bastion host is successfully penetrated the traffic is protected from snooping. To get the internal network and internal traffic, the attacker must penetrate the interior router. With a merged bastion host and

interior router, screened host firewall architecture is present and hence if the bastion host is penetrated, there is no way of security between the bastion host and the internal network.

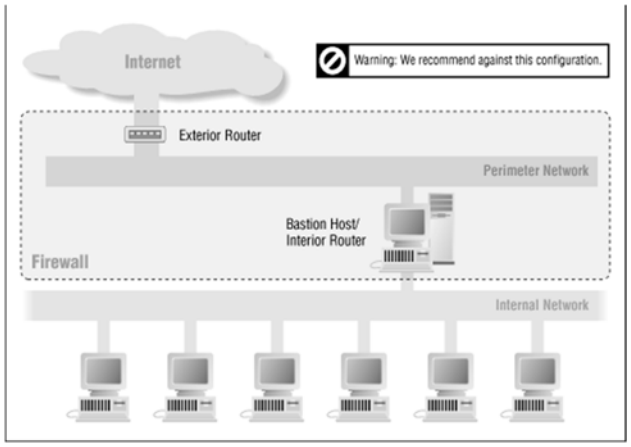


Figure 3.4: A merged bastion host and interior router architecture

One of the main purposes of the perimeter network is to prevent the bastion host from being able to snoop on internal traffic. Moving the bastion host to the interior router makes all of the internal traffic visible to it.

3.5 Dangers of using Multiple Interior Routers

Using multiple interior routers to connect the perimeter net to multiple parts of the internal net can cause many security problems. The basic problem is that the routing software on an internal net part could decide that the shorter route to another internal net part is via the perimeter net and may route it accordingly. Sometimes this traffic may be blocked by the packet filtering on one of the routers and sometimes, if it works, then it provides sensitive, strictly internal traffic flowing across the perimeter net, where it can be snooped on if somebody has managed to break into the bastion host.

It's also difficult to keep multiple interior routers correctly configured. The interior router is defined with the most important and complex set of packet filters and having two of them doubles the chances of getting the rule sets wrong.

On a large internal network, having a single interior router may not be effective in both performance and reliability wise and it may lead to single point of failure that is a major annoyance in providing redundancy. It is safe to set up each interior router to a separate perimeter net and exterior router. This configuration increases performance and redundancy however it is more complex and expensive. It is highly unlikely that traffic will try to go between the interior routers (if the Internet is the shortest route between two parts of the internal network, it can cause much worse problems than most sites) and extraordinarily unlikely that it will succeed.

If there exists any performance problems one may be motivated to look at multiple interior routers, then it's difficult to justify the separate perimeter networks and exterior routers. In the following cases the interior router is the performance bottleneck. One is, a lot of traffic going to the perimeter net that is not then going to the external network. Another is exterior gateway is much faster than the interior gateway. In the first case, because of misconfiguration the perimeter net may take the occasional traffic (that is not significant)that is'nt destined for the external world in some configurations for example, DNS queries about external hosts when the information is cached. In the other case ,instead of adding a second one upgrading the interior router should be considered seriously to match the exterior router. Figure 3.5 shows the basic architecture using multiple interior routers.

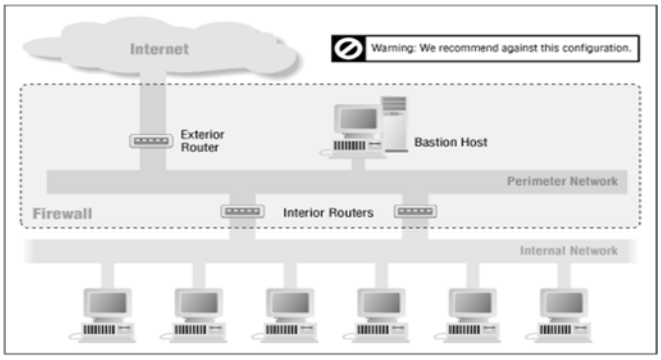


Figure 3.5: Architecture using multiple interior routers

Another reason for having multiple interior routers is to have multiple internal networks, which have many reasons not to share a single router. Giving separate interfaces to a single router can be the simplest way to accommodate these networks as shown in Figure 3.6 .Though this complicates the router configuration it doesn't produce the risks of a multiple interior router configuration. In the case where there are too many networks for a single router or when sharing the router is unacceptable due to any reason, making an internal backbone and connecting it to the perimeter network with a single router, is to be considered as shown in Figure 3.7 .

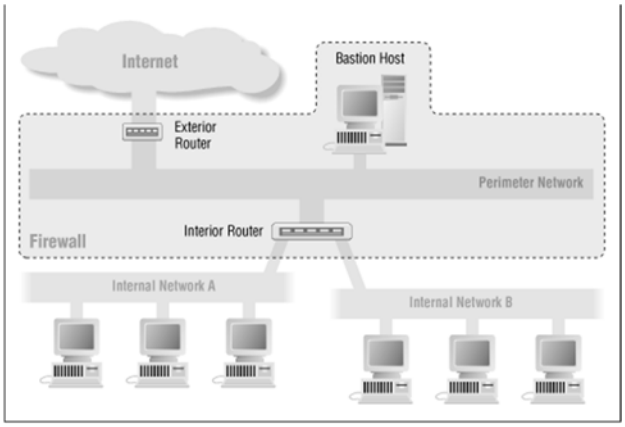


Figure 3.6: Multiple internal networks (separate interfaces in a single router)

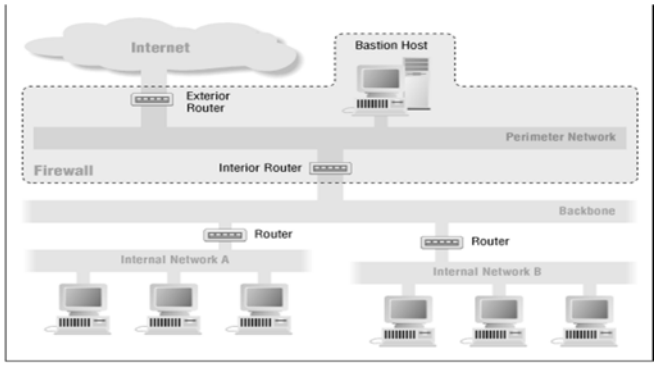


Figure 3.7: Multiple internal networks (backbone architecture)

When one network wants to allow connections that others consider insecure, then an effective way to accommodate different security policies among different internal networks is to attach them to the perimeter through separate routers, In such case, the perimeter network should be the only interconnection between the internal networks; there should be no confidential traffic passing between them; and each internal network should treat the other as an untrusted, external network. Though this is likely to be extremely inconvenient for some users on each network, yet anything else will either compromise the security of the site as a whole or remove the distinction that caused to set up the two routers in the first place.

3.6 Usage of Multiple Exterior Routers

Cases where multiple exterior routers are connected to the same perimeter net, are shown in Figure 3.8 . Examples are

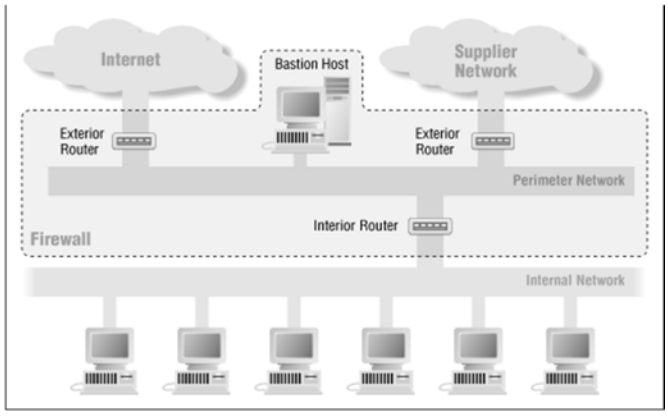


Figure 3.8: Architecture using multiple exterior routers

One is having multiple connections to the Internet for example, through different service providers, for redundancy. Another one is having a connection to the Internet and additional other connections to other sites. In such cases, there might be one exterior router with multiple exterior network interfaces. It doesn't pose a significant security problem by attaching multiple exterior routers which go to the same external network (e.g., two different Internet providers). Though they may have different filter sets, yet that's not critical in exterior routers. Though there is an increased risk of compromise, yet such a compromise of an exterior router is not threatening.

Things are more complex if the connections are to different places (for example, one is to the Internet and one is to a site you're collaborating with and need more bandwidth to). To figure out whether such an architecture makes sense in these cases,this question is to be asked: what traffic could someone see if they broke into a bastion host on this perimeter net? For example, if an attacker broke in, could he snoop on sensitive traffic between the intended site and a subsidiary or affiliate? If so, then installing multiple perimeter nets instead of multiple exterior routers on a single perimeter net, may be considered.

3.7 Multiple Perimeter Networks

In certain situations configuration may include multiple perimeter networks. Figure 3.9 shows this configuration.

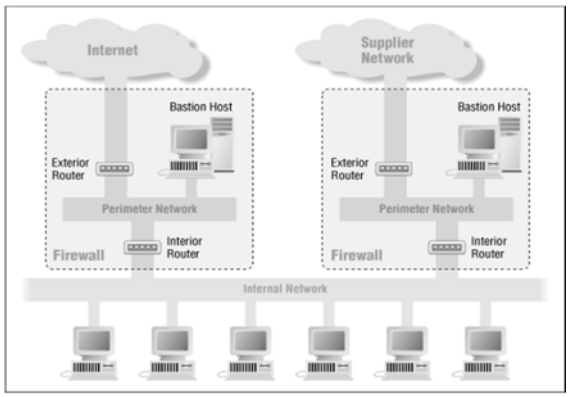


Figure 3.9: Architecture using multiple perimeter nets (multiple firewalls)

Redundancy might be provided by multiple perimeter networks. It is not preferable to invest for two connections to the Internet, and then run them both through the same router or routers. Putting in two exterior routers, two perimeter nets, and two interior routers ensures that there is no single point of failure, Provided, of course, that two Internet providers are actually running on different pieces of cable, in different conduits. Also the destructive power of a backhoe should never be underestimated.

Having multiple perimeter nets is less risky than having multiple interior routers sharing the same internal net, but it still suffers a maintenance problem. Multiple interior routers

may present multiple possible points of compromise. Those routers must be watched very carefully to keep them enforce appropriate security policies; if they both connect to the Internet, they need to enforce the same policy.

3.8 Usage of Dual-Homed Hosts and Screened Subnets

Significant increase in security can be obtained by combining a dual-homed host architecture with a screened subnet architecture. To construct this, the perimeter network has to be split and dual-homed host is to be inserted. The routers provide protection from forgery, and protect from failures where as the dual-homed host begins to route the traffic. The dual-homed host provides finer controls on the connections than packet filtering. This is a belt-and-suspenders firewall, providing excellent multilayered protection, although it requires careful configuration on the dual-homed host to ensure taking full advantage of the possibilities.

4. BASTION HOST GENERIC ARCHITECTURE

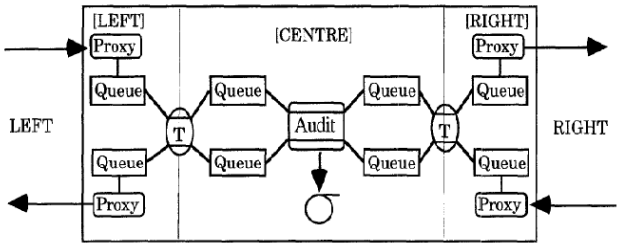


Fig 4.1 The internal Architecture of Bastion host

The Compartmented Mode Workstation (CMW) in Figure is divided into three sections LEFT,CENTRE and RIGHT. Processes serving the LEFT half of the CMW are labelled with a Sensitivity Label (SL) including a category of LEFT, where as processes serving the RIGHT half of the CMW are labelled with an SL including a category of RIGHT.

The CMW is a dual-ported machine and it is configured so that all data coming from the LEFT system is labelled with an **SL** including the category LEFT and all data coming from the RIGHT system is labelled with an SL including the category RIGHT. The networking of the CMW must be evaluated for the sake of assurance, so that confidence can be placed in the data arriving at the network interfaces being labelled correctly.

It is to be noted that the protocols used between the connected systems and the CMW bastion host is unlabelled Internet Protocol (IP) and not Trusted Systems Information exchange (TSIX).LEFT, RIGHT and CENTRE are arbitrarily chosen disjoint categories and are not related in any way to labels that may exist in either end system. The advantage of the disjoint SLs is that unprivileged processes (i.e. the proxies) in one section of the CMW cannot read, move or write data to the other half. the operation of the bastion host and how a message passes from the LEFT system to the RIGHT system is shown in the fig 4.1. The top most proxy in the LEFT section handles the initial connection request.This proxy is unprivileged and is constrained by the CMW access controls to write data only into the queue directory shown.

A trusted (i.e. security critical and evaluated), privileged process, marked as **T** in the figure, moves the contents of the queue across the CMW labelling boundary into the upper left queue in the CENTRE section of the CMW[2].Within the CENTRE section of the CMW, an audit process operates. This process, which must be evaluated to assure that it correctly records the necessary accounting information regardless of any input it receives (i.e. it is not possible to overrun the process). The audit process makes a copy of the message, and moves the message to the upper right queue in the CENTRE

5. BASTION HOST HARDENING

Hardening is the process of configuring a machine to be especially secure and resistant to attack .

The basic hardening process is as follows:

1. Secure the machine.
2. Disable all non required services.
3. Install or modify the services to be provided.
4. Reconfigure the machine from a configuration suitable for development into its final running state.
5. Run a security audit to establish a baseline.
6. Connect the machine to the network it will be used on.

Caution should be taken to ensure that the machine is not accessible from the Internet until the last step. If the intended site isn't yet connected to the Internet, turning on the Internet connection can be avoided until the bastion host is fully configured. If a firewall is to be added to a site that's already connected to the Internet, the bastion host need to be configured as a standalone machine, unconnected to the network.

If the bastion host is vulnerable to the Internet while it is being built, it may become an attack mechanism instead of a defense mechanism. An intruder who gets in before the baseline audit is run, will be difficult to detect and will be well positioned to read all the traffic to and from the Internet. Cases have been reported where machines have been broken into within minutes of first being connected to the Internet; while rare, it can happen.

Copious notes need to be taken on every stage of building the system. Assume that sometime in the future, a compromise may cause the machine to burst into flames and be destroyed. In order to rebuild the system, all the steps that were taken previously need to be followed.As all of the software that was used is required, it is to be ensured to securely store all the things needed to do the installation, including:

- The disks, CDs, or tapes to install software from
- The source code for any software which is built from the source
- The environment used to build software from [3]
- source, if it's different from the one that is being installed; this includes the operating system, compiler, and header files (and a machine they run on)

5.1. PRECAUTIONS FOR HARDENING THE BASTION HOST

Due to the level of required access with the outside world, bastion hosts are particularly vulnerable to attack. To make the bastion host hardened, the following precautions need to be followed.

Disable or remove any unnecessary services or daemons on the host. Disable or remove any unnecessary user accounts. Disable or remove any unnecessary network protocols. Use encryption and multi-factor authentication for logging into the server. Configure logging and check the logs for any possible attacks. Run an intrusion detection system on the host. Patching the operating system with the latest security updates. Lock down user accounts as much as possible, especially root or administrator accounts. Close all ports that are not needed or not used.

.

7.CONCLUSION & FUTURE SCOPE:

Bastion Hosts sitting on the network perimeter are designed for secure information flow between public network and private network. There may be several roles of Bastion hosts like router, DNS, FTP, SMTP, News, and/or Web servers. The role of System administrator is to identify the services needed on Bastion host to resist the possible attacks. The hardening of Bastion hosts allows to resist attacks from external sources thus protecting the internal network. The different bastion host architectures and hardening methods are discussed. This work can be extended to propose a new method for effective utilization of services using priority queue for the required services. It can be further extended to use logs to prioritize the services on the multi homed multiple bastion hosts.

References

[11 Chapman, D. Brentand Elizabeth D. Zwicky., Building Internet Firewalls, Sebastopol CA OReilly & Associates 2000

[2] Chris Cant & Simon Wiseman., Simple Assured Bastion Hosts Computer Security Applications Conference, 1997. Proceedings., 13th Annual

[3] Richard E. Smith., Mandatory Protection for-Internet Server Software, Procs. 12th Annual Computer Security Applications Conference, San Diego, December 1996, pp. 178-184.

[4] R. Knobbe; A. Purtell; S. Schwab., Advanced security proxies: an architecture and implementation for high-performance network firewalls ,DARPA Information Survivability Conference and Exposition, 2000. DISCEX '00. Proceedings Year: 2000, Volume: 1 pp. 140 – 148

[5] Lodin, S., and Schuba, C.,Firewalls Fend Off Invasions from the Net. IEEE Spectrum, February 1998

[6] Oppliger, R., Internet Security: Firewalls and Beyond Communications of the ACM, May 1997.

[7] Mohammed G.Gouda;Alex X.Liu.,Structured firewall design ,Computer Networks: The International Journal of Computer and Telecommunications Networking archive Volume 51 Issue 4, March, 2007 pp.1106-1120

A unique dimensionality shrinkstylefor high-dimensional spatiotemporal brain signal data based on graph signal processing theory

Dr.D.Nagaraju[1],
Professor,Dept of Information Technology,
LakireddyBalireddy College of Engineering,
dnagaraj_dnr@yahoo.co.in

A.Sarvani[2],
Asst.Professor,Deptof Information Technology,
LakireddyBalireddy College of Engineering,
sarvani.anandarao@gmail.com

B.Venugopal[3],
Asst.Professor,Dept of Computer science engineering,
Andhra Loyola Institute of Engineering and Technology,
srees.boppana@gmail.com

ABSTRAT

EEG or MEGis brain imaging data which is high-dimensional spatio temporal values that frequently need dimensionality shrink. Now this low dimensionality data which we got after application of dimensionality shrink can be used in multiple applications. This paper showboat a unique dimensionality shrink style. To get the reliable measurements here we have used unique graph signal processing model. Here the main goal is tosegregate the brain imaging signals which we got asoutcome to visual stimuli. Here to develop a connectivity graph we utilize relaxing state measurements of the subjects. To build a low-dimensional linear subspace for the task-state measurements we use the graph Laplacian and Graph-Based Filtering (GBF) Here we develop the connectivity graph appropriate for this application using numerous means. Whennon-Gaussian noise corrupted the measurements then we can use the connectivity graph information to get the accurate low dimensional data

Keywords— Brain imaging, Dimension shrink, Graphbased filtering, Graph signal processing

I. INTRODUCTION

In neuroscience and brain computer interface (BCI) technology, automaticscrutiny of brain imaging data plays a key role. The job is to discover the spatiotemporal neural signature of a task, which can be done by implementing segregation on cortical activations which arouse by diverse stimuli [1, 2]. Universal brain imaging means are Electroencephalography (EEG) and Magnetoencephalography (MEG). MEG evaluates the magnetic fields generated bymovement of electrical signals in the brain using highly responsive sensors placed across the scalp. Then we get these measurements in high dimensional spatiotemporal data. In our experiments, to record the brain electrical signals for the time span of 1100 milliseconds we used recumbent Elekta MEG scanner which are having 306 sensors. Here we use highly responsive sensors to record the brain stimuli .Further high, the measurements deteriorating by numerous types of noise (e.g., sensor noise, ambient magnetic field noise, etc.). Due to high-dimensionality and noise, the accuracy and speed of the signal scrutiny gets reduced. But this produces inaccurate signature modelling for segregate. Due high-dimensionality of these signals, complexity of classifier also rises. It is very difficult to develop a model with both complex classifier and availability of few data samples. Thus to bring out robust dimensionality shrinkstyle we require solid study of brain imaging data. Using this we can easily insert of task-related information .This insertion of task related information is called transformation process.

We can get low-dimensionality data using dimensionality shrinkmeanstransforms with a linear or nonlinear style by mapping the high dimensional data on to the space of low Linear discriminate analysis (LDA).The following comes under the linear styles of dimension shrink methods: locality preserving projections (LPP), marginal Fisher analysis (MFA).We can also use multiple nonlinear styles for dimensionality shrink. The following comes under the non-linear styles of dimension shrink methods: Self-organizing maps and other

neural network-based approaches (e.g., autoencoder [4]).To build the geometric structure of manifold we can use Laplacian eigenmaps (LE) [5] and diffusion maps [6].This geometric structure helps to map the data points into a lower dimensional space

In this study, we bring out two things one is unique graph signal processing theory [7] and other is graph based filtering algorithm (GBF).These two above approaches helps in performing dimensionalityshrink,andto build a connectivitygraphwhich is best suitablefobrainimaging. GBF-based model can reduce the dimensionality in a high robust manner as it uses the underlying graph model. This graph model can be used as side information to bring out accurate data. This above approach does not exists in traditional dimensionality shrinkapproaches. This side information (modelled as graph) can be used to produce accurate data as thiscontains high reliable datathan measurements, when the measurements are depraved by noisy. We can get high reliable low-dimensional subspace by investing on the graph. In the previous work to build low dimensional subspace/manifold we use solely measurements. (Eg: PCA, Laplacian Eigen maps) but does not depend on graph model

Therefore, here first we collect relaxing-state brain imaging signals and then on that we apply connectivity scrutiny on thesignals. Then, after we got relaxing-state connectivity graphweapply Laplacian on the graph, and then by applying dominant eigenvectors we develop low-dimensional subspace. Next, we take low-dimensional subspaceto map the noisy task-state measurements (which arouse by visual stimuli) which leads to dimensionality reduction signals called as low-dimensional signal. Then on this reduced-dimensional signal we apply SVM classifier .If we apply SVM classifier then we can getinsertion of task-related discriminative information

GBF used a “normal” linear subspace for detecting abnormality in actuator networks in the paper written by [8].In [8], we are said to have abnormal measurements when they vary from the normal subspace. In image compression [9] and temperature data [10] GBF has given the accurate data than any other technique. In [11, 12] some signal features are recommended which we can get from graph signal processing. In [13], brain signals are classified into small, median, large frequency components using graph Fourier transform. This information is used for scrutiny on properties of functional brain connectivity. The main aim of our work is to concentrates on linear dimensionality shrink .Here we also use graph Laplacian eigenvectors to attain the low-dimensional data. Here to explore the usefulness of this approach, on brain image signalswe use classification task. There is vast difference between the recommended GBF-based approach and Laplacian eigenmaps [5].To get low-dimensional manifold we need Laplacian eigenmaps. This Laplacian eigenmaps consider only the measurement. In Laplacian eigenmaps the side information is not represented. The production of low-dimensional data uses a very unique mechanism

II. GBF BASEDDIMENSIONALITYSHRINK:

For a reliable dimensionality shrinkwe take the help of GBF which can be done in three levels: First level is pre-processing, second level is decomposition and third level is projection. In the first level, by considering the application or data we compose the graph G (N, ε, A), Here N points to the nodes or sensors in the graph and ε pointsto connecting edges of the nodes. Then we can calculate the adjacency matrix W using the graph G. We can calculate edge weights between twodistinct nodes by taking the help of adjacency matrix. W can be calculated by many ways. One way to calculate W is :

$$A_{ij} = \exp\left(-\frac{(1-\|\rho(i,j)\|1)^2}{2\sigma_1^2}\right).\exp\left(-\frac{d(i,j)^2}{2\sigma_2^2}\right) \quad (1)$$

Here the correlation coefficient of the covariance matrix Cwith elements is represented by $\rho(i,j)=c_{i,j}/(\sqrt{(c_{i,i}c_{j,j})})$.The Euclideand distancebetweensensoriand j which is normalized is represented by $d(i,j) \in [0,1]$. Highdecayrates are represented by σ_1 andσ_2. In the secondlevel, byusing the formula $L = I - D^{-1/2} WD^{-1/2}$, we acquire the normalized Laplacian matrix L.In the formula D is equal to diagonal degree matrix. This can be calculated by using the formula$D_{ii} = \sum_j w_{ij}$. Then we solve L using the formula $L = U\Lambda U^T$ where U is the eigenvectors $\{u_l\}_{l=1,\ldots,N}$matrix. We sort the Eigen values from higher to lower to acquire the terminationfrequency. Subspace can be constructed by the eigenvectors whose

eigenvaluesare greater than termination frequency. In the third level, we use the constructed subspace to get the low-dimensional data by mapping the high dimensional data i.e. original data on to the subspace.

III. EVALUATION ON SYNTHETIC DATA

In this, we can perform comparison between GBF, PCA and LDA by applying dimensionality shrink method on noisy synthetic data. Then for GBF, PCA and LDA we implement the binary SVM classifier individually, and to get robust method, the outcome is compared by applying individual noise. Finally after completion of the entire process we can guarantee that GBF is the maximum robust method and this GBF can also produce the accurate results in the case when noise is non-Gaussian.

Consider a case where graph consists of m nodes where m=100, connecting probability between the nodes is p where p=0.3. For generating the signal we take the Eigen vectors where k=8 and we call this Eigen vectors as signal generating components. Create the two data class, for first one set β_1 =4, γ_1 =2.5 and for second one set β_2 =2.5, γ_2 =4. Set μ =0 and variance σ =1 /6 for three types of noises. Now our aim is to decrease the dimensionality from m = 100 to 8 by applying GBF, PCA and LDA separately

Table 1:Classification accuracy with synthetic data.

Styles	Gaussian Noise	White Uniform Noise	Sparse White Noise
GBF	0.8564	0.8718	0.8744
PCA	0.8641	0.8564	0.8590
LDA	0.8385	0.8436	0.8333
original	0.8692	0.8718	0.8564

For GBF, we use the parametersσ1 =0 .2 and σ2 =0 .9 to calculate the adjacency matrix W. SVM classifier performance is compared with 10-fold cross validation. Table 1 consists of accuracy of each classification. Here for white noise we take SNR=25dBand for sparse white noise we consider SNR =25dB, e% = 5%.Using the above table, it is clear that in the case of Gaussian noise the PCA gives the better performance. So to reduce the dimensionality when there is Gaussian noise PCA is productive. But in the case of non-Gaussian noise, GBF is productive and provides the accurate results

In this case robustness of GBF is calculated when noise = white noise for different levels of signal to noise ratio i.e. when SNR = 10dB to SNR = 60dB.Here it is proven that GBF is inversely proportional to 1/SNR. When we apply GBF on the original data we can get more accurate and robust measurements. Again performance of GBF is calculated when noise = sparse white noise for different levels of percentage to noise ratio i.e. when SNR = 25dB, e=50% to e=1%.In both the cases GBF has performed better than PCA.Now we do final evaluation with real MEG data after the evaluation with synthetic data is completed.

IV. BRAINCONNECTIVITYSCRUTINY

The most important and general problems in GBF to brain signal processing is to find an accurate connectivity graph. We need complex and sophisticated scrutiny method to gather the brain signals as because our brain is a huge complex network. There are three unique brain connectivity: first one is structural connectivity, second one is functional connectivity and the last one is effective connectivity. GBF graph can be constructed using these three connectivity. This graph is essential in building the side information. The physical connections of the brain help to form structural connectivity. A functional dependency between numerous brain regions gives raise to functional connectivity. For building Effective connectivity we need neuronal system components causal relationship or Directed influence .Here we have used three functional connectivity and one effective connectivity to build GBF graph. In each case, to develop the connectivity graph, we examine the relaxing state recordings (measurements before the onset of the stimulus) and we utilize this relaxing-state connectivity graph to build the task-state measurements (after the onset of the stimulus). If relaxing-state sensor measurements (time-series signals) at two spatially-separated brain regions: X and Y written as x_t and y_t for t =1, 2, ...,T, we discuss how to compute the edge weight $w_{x,y}$.

A. Correlation connectivity

Correlation is a primary estimation for statistical dependency for function connectivity [17]. We consider the correlation coefficient between X and Y as the edge between these two regions:

$$w_{x,y} = \frac{\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y})}{(T-1)s_x s_y}$$

$$\frac{\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2\ \sum_{t=1}^{T}(x_t-\bar{x})^2}} \quad (2)$$

Where$\bar{x}$and $\bar{y}$are the sample mean of X and Y.s_x and s_y are sample standard deviation of X and Y.

B. Coherence connectivity

For oscillatory interdependency connecting the two diverse brain regions first we need to estimate the coherence connectivity. Cross-correlation coefficient depends on frequency domain analog. Consider two signals x (t), y (t) having frequency f, now to calculate transitory phase of the signal with respect to time we should spectrally decompose at threshold f. For each individual signal we apply the band-pass filtering where frequency is equal to f ± 5Hz, the loop of f(t) with a Morlet wavelet centred at frequency f put forward the transitory phase at time t. Here we consider two signals $x = A_x(t)e^{j\varphi x(t)}$ and $y = A_y(t)e^{j\varphi y(t)}$, amplitudes are represented by $A_x(t)$ and $A_y(t)$,phases at time t for two signals are represented by $\varphi_x(t)$ and $\varphi_y(t)$ .Then using the below formula we can calculate coherence connectivity edge:

$$w_{x,y} = \left| \frac{\frac{1}{T}\sum_{t=1}^{T} A_x(t)A_y(t)e^{j[\phi_x(t)-\phi_y(t)]}}{\sqrt{1/T\sum_{t=1}^{T} A_x(t)^2}\cdot\sqrt{1/T\sum_{t=1}^{T} A_y(t)^2}} \right| (3)$$

C. Phase locking value (PLV) connectivity

We use the methods described in [19] [20] on MEG data to get the phase locking value. Here we take time-series signals for a particular frequency to measure phase synchrony which is said to be PLV.If amplitude effects from the consistency of the phase difference there will be deviation of PLV from coherence connectivity [19]. So measuring procedure followed in coherence connectivity in applied here. After we get $x = A_x(t)e^{j\varphi x(t)}$ and $y = A_y(t)e^{j\varphi y(t)}$,here we can calculate time averaged value by measuring the edge connecting the X,Y regions

$$w_{x,y} = \left| \frac{\frac{1}{T}\sum_{t=1}^{T} 1\times 1e^{j[\phi_x(t)-\phi_y(t)]}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} 1^2}\cdot\sqrt{1/T\sum_{t=1}^{T} 1^2}} \right|$$

=

$$\frac{1}{T}\left|\sum_{t=1}^{T} e^{j[\phi_x(t)-\phi_y(t)]}\right| \quad (4)$$

D. Grangercausalityconnectivity

Here we consider two regions in both directions and at numerous frequencies to calculate the causality relationship using Granger Causality. The directed interactions between neural assemblies are said to be calculated using causality relationship. For each individual frequency we calculate the Grangercausality. The ratio of predicted power to total power gives raise to granger causality [21]. First we choose the two regions called X and Y now fix univariate and bivariate AR models in that regions. This AR model consider the past of the signal which is calculated from one region to predict the signal from other region. To get the prediction error using univariate AR model we consideronly past of the own signal [22]:

$$\sum_{k=0}^{m} a_{1k}x^{(t-k)} = u_1(t)$$

$$\sum_{k=0}^{m} b_{1k}y^{(t-k)} = v_1(t) \quad (5)$$

To get the prediction error using bivariate AR model we consider both past of the own signal and past of the other signal [22]

$$\sum_{k=0}^{m} a_{2k}x(t-k) + \sum_{k=1}^{m} c_{2k}y(t-k) = u_2(t)$$

$$\sum_{k=0}^{m} b_{2k}x(t-k) + \sum_{k=1}^{m} d_{2k}y(t-k) = v_2(t) \quad (6)$$

u_1,u_2,v_1,v_2 represent the uncorrelated residual errors and a_1,a_2,b_1,b_2,c_2,d_2represent the model coefficients .Accuracy of prediction can be calculated only if we have variance of the prediction errors

$$\sum\nolimits_{X|X\_} = var(u_1), \sum\nolimits_{Y|Y\_} = var(v_1)$$

$$\sum\nolimits_{X|X\_,Y\_} = var(u_1), \sum\nolimits_{Y|Y\_,X\_} = var(v_1) \quad (7)$$

Consider the two signal X and Y. The Granger causality between those two signals are defined by [23]:

$$GC_{Y\to X} = \ln\frac{\sum_{X|X\_}}{\sum_{X|X\_,Y\_}}$$

$$GC_{X\to Y} = \ln\frac{\sum_{Y|Y\_}}{\sum_{Y|Y\_,X\_}}$$

(8)

Now we take the maximum Granger causality value. This maximum value represent the strength of interaction. Using this maximum value symmetric matrix of GBF is build which is the adjacency matrix.

$$w_{x,y} = \max(GC_{Y\to X}, GC_{X\to Y}) \quad (9)$$

V. EVALUATIONON MEG DATA

Using real MEG data, we check our algorithm. Initially, we check all types of connectivity graphs. By using relaxing state measurements i.e. 100ms before the stimuli appears on the real data the graphs can be generated. To evaluate subsequent experiments we use the value with best performance. We measure dimension shrink performance for diverse algorithms and then we compare. Here we have performed comparison between GBF, PCA, LDA, LE algorithms. To acquire and measure connectivity graphs here we have used MATLAB open source toolbox BrainStorm [24].

Table 2:Segragate accuracy with MEG data

Style	Accuracy	Data Dimension
GBF(GCC)	0.6800	21
PCA	0.6688	21
LDA	0.6690	1
LE	0.6296	21
Original	0.6708	306

A. *Comparison between diverse graphs*

First we show stimuli to the person then we acquire the brain signals from 96ms to 116ms time interval. The interval from 96ms to 116ms contains maximum discriminative information and this time interval is called cortical activities [2].To get the reliable calculation we have taken 20ms-long sliding window. Then perform average on the data which gives 306 vector. These 306 vector are called dimension vector. Here we taken Step length is 1ms.

The GBF combined with Granger Causality connectivity (GCC) gives the outstanding performance. Therefore, in our experiment we used the granger causality connectivity

B. *Comparison of GBF,PCA and LDA*

Here we have taken step length to 1ms and 20ms-long window in our experiment. To shrink the dimension we have used GBF, PCA, LDA and LE. We used SVM classifier for categorization. Now to get the final output median the accuracy with respect to time. For documenting the median classification accuracy, we take 10-fold cross validation.

Here the high dimensionality is reduced to low dimensionality. i.e. from 306 dimensions to 21 dimensions using both GBF and PCA. But GBF have produced better performance than PCA. This can be easily identified by the above table.

VI. EXPERIMENT

In our experiment we have taken two types of data one is synthetic data and other is real MEG data. Then we have applied many dimensionality shrink technique on both synthetic data and also on real MEG data by this we can easily find out which technique produces the best result. We generate spatiotemporal signals by combining Gaussian and non-Gaussian noise in the case of synthetic data. The sum of both signal S_t and noise N_t gives synthetic signal H_t

$$H_t = S_t + N_t \quad (10)$$

S_t points to signal vector and N_t points to noise vector,

S_t can be acquired from graph information. The graph consists of m nodes and we form many pair of nodes having the probability of p. The range of edge weights vary is between 0 to 1 with uniform distribution between two diverse nodes. Here we have used m×m adjacency matrix W which is symmetric to represent the above distribution. Then we use Laplacian matrix L to do the decomposition. This decomposition performed by Laplacian matrix give eigenvectors.

$$S_t = \beta \sum_{i=1}^{\frac{k}{2}} a_{t,i} f_i + \gamma \sum_{i=\lceil \frac{k}{2} \rceil + 1}^{k} a_{t,i} f_i \quad (11)$$

Components scalar weight is represented by β and γ .Here we have Laplacian matrix Lconsists ofeigen vectors where i-th eigenvector is represented byf_i.To generate signal we use 'k' number of eigen vectors, and we select some uniform random variable represented as $a_{t,i}(1 \leq i \leq k)$ in the range [0, 1] [14].By changing β and γ values we can produce manydiverse classes. These diverse classes are used for binary classification

Here we use three types of noises, one is white noise with Gaussian distribution, second one is white noise with uniform distribution, and third one is sparse white noise. These noise components are combined with signal to generate synthetic signal. We choose e% nodes from total number of nodes to produce sparse white noise, this e is some random number. Now we add this noise to our signal which produce final high dimensional data H_t. In this way we calculate n temporal samples. The final data consists of: $[H_1, H_2, ...H_n]$.

After we acquire the synthetic data we move to second part of an experiment, in this part MEG signals are segregated. To record the MEG signals we show many objects and faces to the person and then we record the response to these faces and objects. These faces and objects are called visual stimuli. In this experiment we used 306 sensors, and showed 320 face images and 192 non-face images .We start this recording from 100ms before the images are shown and continued till 1 second after the images are shown (i.e. 1100ms in total).The image i.e. stimuli is placed in the centre of the screen for 300ms and person is asked to concentrate on each image. Here we don't take any behavioural responses of the person, we consider only the brain signals. Now to find the best technique in dimensionality reduction. We apply PCA, LDA, LE and GBF on both synthetic data and also on MEG data. Nowto segregate based on the stimuli viewed we have used SVM. Based on segregate, performance can be calculated.

VII. CONCLUSION

When we have a data which can be represented by a graph the best technique is Graph-based filtering (GBF).This is one of the signal processing technique. This paper has brought dimensionality shrinkingofbrain imaging data using GBF. The GBF has produced reliable and accurate data than other techniques in the context of dimensionality shrink. This is proven from outcomes we got by using both synthetic data and MEG data, GBF produces the excellent performance when signal hasnon-Gaussian noise, generally brain imaging details in this form. Our future scope is to solve the complexity of GBF and its usage in real-time applications

REFERENCES

[1]Thanou, Ntorina. "Graph Signal Processing." (2016).

[2] Fischl B, Dale AM. Measuring the thickness of the human cerebral cortex from magnetic resonance images. ProcNatlAcadSci U S A. 2000;97:11050-11055.

[3] Shuman, David I., et al. "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains." IEEE Signal Processing Magazine 30.3 (2013): 83-98.

[4] Sui, Jing, et al. "A review of multivariate methods for multimodal fusion of brain imaging data." Journal of neuroscience methods 204.1 (2012): 68-81.

[5] Mikhail Belkin and ParthanNiyogi, "Laplacian eigenmaps for dimensionality shrink and data representation," Neural computation, vol. 15, no. 6, pp. 1373–1396, 2003.

[6] Rui, Liu, HosseinNejati, and Ngai-Man Cheung. "Dimensionality reduction of brain imaging data using graph signal processing." 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016.

[7] KleovoulosTsourides, ShahriarShariat, HosseinNejati, Tapan K Gandhi, Annie Cardinaux, Christopher T Simons, Ngai-Man Cheung, Vladimir Pavlovic, and PawanSinha, "Neuralcorrelatesofthefood/non-

foodvisualdistinction," Biological Psychology, 2016

[8] Jieqi Kang, Shan Lu, Weibo Gong, and Patrick A Kelly, "A complex network based feature extraction for image retrieval," in Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014, pp. 2051–2055..

[9] Richiardi, J., Achard, S., Bunke, H., & Van De Ville, D. (2013). Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience. IEEE Signal Processing Magazine, 30(3), 58-70.

[10] Fallani, F. D. V., Richiardi, J., Chavez, M., &Achard, S. (2014). Graph analysis of functional brain networks: practical issues in translational neuroscience. Phil. Trans. R. Soc. B, 369(1653), 20130521.

[11] Xiaowen Dong, Antonio Ortega, Pascal Frossard, and Pierre Vandergheynst, "Inference of mobility patterns via spectral graph wavelets," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 3118–3122.

[12] Kozma, Robert, and Marko Puljic. "Random graph theory and neuropercolation for modeling brain oscillations at criticality." Current opinion in neurobiology 31 (2015): 181-188.

[13] Xudong Jiang, "Linear subspace learning-based dimensionality shrink," Signal Processing Magazine, IEEE,vol.28,no. 2, pp. 16–26, 2011.

[14] Kim, W. H., Adluru, N., Chung, M. K., Okonkwo, O. C., Johnson, S. C., Bendlin, B. B., & Singh, V. (2015). Multi-resolution statistical analysis of brain connectivity graphs in preclinical Alzheimer's disease. NeuroImage,118, 103-117..

[15] Hugdahl, K., &Westerhausen, R. (2015). Speech processing asymmetry revealed by dichotic listening and functional brain imaging.Neuropsychologia..

[16] Ed Bullhigh and Olaf Sporns, "Complex brain networks: graph theoretical scrutiny of structural and functional systems," Nature Reviews Neuroscience, vol. 10, no. 3, pp. 186–198, 2009.

Support Vector Machine Based Classification for Face Recognition

D Sudha Rani[#1] ,P Swathi[#2], V Srinivasa Rao[#3], K Srinivas[#4]

[#]*Department of Computer Science and Engineering , VRSiddhartha Engineering College, Vijayawada, Andhra Pradesh, India.*

[1] devarapalli.sudharani@gmail.com
[2]swathi.pappusetty@gmail.com
[3]drvsrao9@gmail.com
[4]vrdrks@gmail.com

Abstract **— Face recognition is an important research field of pattern recognition. Up to now, the face recognition caused many researchers great concern from the fields such as pattern recognition and computer vision. In general, we can make sure that the performance of face recognition system is determined by how to extract feature vector correctly. This paper presents a face recognition method based on Gabor filter and Support vector machine. At first, the Gabor filter bank with four frequencies and eight orientations is applied on each face image to extract the features against local distortions of facial expression. Finally, Support Vector machine is used for classification of the extracted features. The present method is tested on the Olivetti Research Laboratory face database.**

Keywords — Face Detection, Feature Extraction, Face Recognition.

I. INTRODUCTION

Support Vector machines are the supervised learning algorithms for both classification and pattern recognition based on learning theory. The basic principle of Support vector machine (SVM) is that to construct a hyper plane as a decision plane to separate the data points between two classes. The data points that are at a distance exactly equal to the hyper plane are called support vectors. SVM is one of the most useful techniques in classification problems. The best example of SVM is face recognition.

SVM is a two class problem. The two classes are

(i) The first one is dissimilarities between faces of different people and

(ii) The second one is dissimilarities between faces of the same person. The motivation of SVM is that to which class the data point will be in.

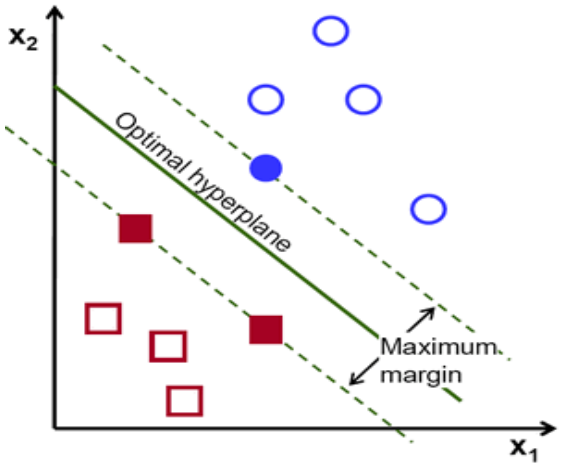


Fig 1: Support vector machine separating the hyper plane of two classes.

The applications of SVM are

(i) Text characterisation.
(ii) Hand –written characterisation.
(iii) Face recognition.
(iv) Image classification.

A. OVERVIEW OF THE PROPOSED METHOD

The flow chart of the proposed method is shown below. First taking the input image from the digital camera and the images are saved in a .bmp format with required size. By using frame differencing the background of the image is eliminated after comparing the two images. The face region is found after matching the binary image and the template image. If the face region is found then the bounding box is appeared on the binary image. After the face region found then the

features are extracted by using Haar Gabor wavelet filters. Face recognition is done by Support vector machine.

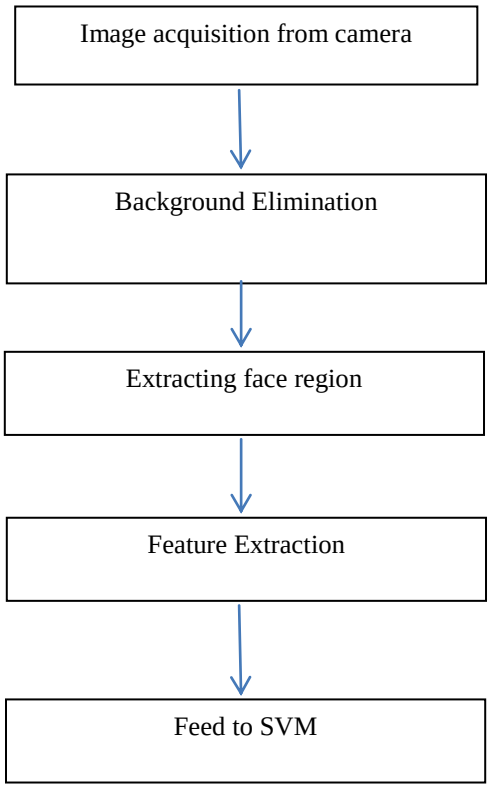


Fig 2: Flow chart of the process.

II. LITERATURE SURVEY

- Face recognition using Gabor filter bank, kernel principal component analysis and support vector machine. In this method the features are extracted by using gabor filter bank and the kernel PCA is performed on the outputs of the filter bank to form the low dimensional feature vectors. Finally SVM is used for classification.
- Face recognition based on SVM and Gabor filter. In this method SVM based face recognition system is proposed. Gabor filters are used to extract the features from the face images. Homomorphic filtering is used for pre-processing before extracting the features. The extracted features are given input to the SVM for training and testing purpose. The difference of this method while compared to other methods is the Yale face database-B is used.
- Face recognition using Gabor wavelet features with PCA and KPCA. In this method Gabor wavelet transform is used and provide a study in depth comparison between Gabor PCA(linear) and Gabor KPCA(non-linear).
- The database used is the ORL database and demonstrated that the Gabor PCA is more efficient than the Gabor KPCA for face recognition system.
- Support vector machine applied to face recognition. In this method the algorithm is applied on both verification and identification and compare the performance of SVM algorithm to a PCA-algorithm. For verification the error rate of SVM was half that of PCA. This indicates that SVM gives more accurate information in the face space.

  The advantage of SVM over other methods like neural networks is the generalisation ability and it gives high accurate results.

III. METHODOLOGY

Background Elimination: We are using frame differencing to eliminate the background elimination. We will take the pixel values of the face image and subtract it with the corresponding pixel values of the background image. Then we get the pixel values of the face image without background.

Face Detection : Template Matching is used for face detection. This approach is to compute the correlation between the templates and each of the known images. We generate four templates of different sizes to test the given images for finding the face region. We run each and every template over a given image and calculate the Euclidian distances between template and face image. The minimum distance template gives the face region.

Feature Extraction : Haar wavelet gabor filters are used to extract the features. The Gabor 2D filter equation is applied and multiplied with the image pixels. This method is worked only on the frontal faces. We calculate the average of each face image per each pixel of different faces. After calculating the mean of each image pixel storing those mean

values in a big array. Then we get the feature vectors.

Face recognition: Support vector machine classifier is used for face recognition. The features that are extracted are given as an input to the SVM. After taking the features of different faces then a new face feature vector is taken and compared the new face image feature with all the face features. Then the new face is generated with those features.

IV. ALGORITHM

Step 1: The input image is taken from the Olivetti face database.

Step 2: The background is eliminated by using frame differencing.

(a) The pixel values of corresponding image are taken in one array.
(b) The pixel values of background image are stored in another array.
(c) Then the two image pixel values are subtracted.
Diff = (a – b) /2
(d) Then we get the face image pixel values without the background.

Step 3: The face region is detected by using template matching.

(a) First the bitmap image is taken and convert the bitmap image into binary image.
(b) After that the 5 templates is taken of different sizes.
(c) Then run each and every template on the corresponding binary image.
(d) Calculate the Euclidian distance between the template and the corresponding image pixel values.
$$\mathbf{D = Sqrt(X_i - X_j)^2 + (Y_i - Y_j)^2}$$
(e) When we get the minimum distance then the template is matched and draw a rectangle box to detect the face region.

Step 4: To extract the features from the frontal faces the haar wavelet Gabor filters is used.

(a) First the bitmap image is converted into gray scale image and
(b) The pixel values of the face database are stored in two different arrays.
(c) For those two arrays sum and difference is calculated.
$$\mathbf{Sum = (a + b)/2}$$
$$\mathbf{Diff = (a - b)/2}$$
(d) After that those values are read onto the image and the image is saved as .bmp format.
(e) That image is divided into four blocks. Two blocks are sum pixel values of which face image is displayed.
(f) Next two blocks is the difference of which black region of some features are displayed and save the image.
(g) Next, that image is taken as input for that four blocks mean and standard deviation is calculated.
$$\mathbf{Mean = \sum X / N}$$
$$\mathbf{Sd = \sqrt{\sum (X - mean)^2 / N}}$$
(h) Like that we take 6 Olivetti face datasets. For each dataset we get 8 values. Those are the feature vectors.

Step 5: Next, the face dataset among the six faces one different pose face dataset is taken and as the same above the feature vectors are generated.

Step 6: Each face feature vectors are subtracted with the corresponding image features.

$$\mathbf{Diff = (a - b) /2}$$

Step 7: The corresponding image features should match with the same face image features.

Step 8: Then the process is positive otherwise negative.

V. ADVANTAGES OF SVM

1. SVM effective in high dimensional spaces.

2. Can improve the accuracy and reduce the computation.

3. Different kernel functions can be specified for the decision function.

4. But it is also caused to specify custom kernels.

IV. RESULTS

The effectiveness of the proposed face detection and recognition algorithms are demonstrated using c#. The face database consists 50 images. First we train the 5 frontal face image features and then we train a new face features and match with all face features. Support vector machine recognises 60% of the faces.

Fig 3: Input images to extract the features

The above face dataset is the Olivetti face dataset. From these face images we calculate the sum and difference based on pixel values of all the images.

Fig 4: Images after calculating sum and difference

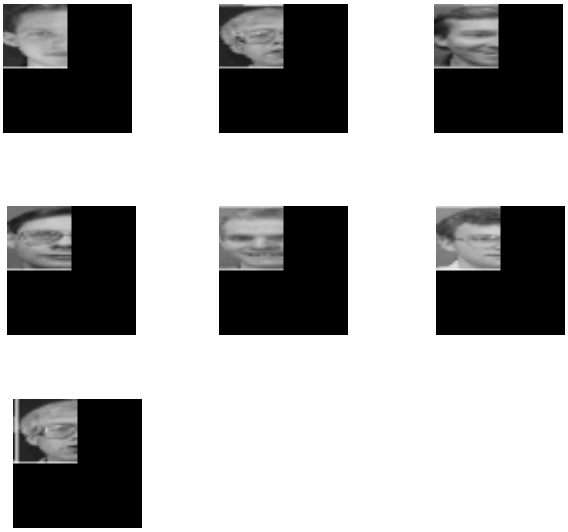


Fig 5: Images after calculating mean and standard deviation

VII CONCLUSIONAND FUTURE WORK

In this paper, we use the Support vector machine classifier and recall technique is proposed for face recognition. Image features are stored in a file and new image feature is taken to test with all the face features. From these results, we can conclude that recognition accuracy is high, very fast and simple. In this paper we have used linear SVM to detect the face region for both face and non-face dataset.

The work further extended by providing the Olivetti face datasets with different poses can be recognized by using Multi-class non-linear SVM and it also helps in increasing the system performance.

VIII REFERENCES

1. Saeed Meshginitober, Ali Aghagolzadeh, “Face recognition using Gabor filter bank,Kernel principal component analysis and Support vector machine, Internatinal Journal Of Computer theory and engineering, Vol.4,No.5ty October 2015.

2. Shruti Y. Bhrind and V.V. Gohokar, “Face recognition based on SVM and gabor filter”, Internatinal Journal of current engineering and technology, E-ISSN-2247-4106, P-ISSN 2347-516.
3. PraseedaLakshmi.V, Dr.M.Sasikumin, ”Analysis of Facial expression using Gabor and SVM”, International Journal of recent trends in engineering, Vol.1, no.2, May 2009.
4. Tudor BARBU, “Gabor filter based face recognition technique”, Proceeding of the Romanian Academy, Series A, Vol.11, Number 3/2010,pg:277-283.
5. P. Jonathon Philips, “Support vector machine applied to face recognition”, National Institute of Standards and Technology.
6. Ashwin Swaminathan, “Face recognition using support vector machines”.
7. Gulonder Aydin Kayaick,”Multi-view face detection using Gabor filters and Support vector machine”, Technical report, IDE0852, May 2008.
8. Navin Prakash, Dr Yashpal Singh, “Support vector for face recognition”, International research journal of engineering and technology, Vol.2, Issue.8,November 2015.
9. Bernd Heisele, Purdy Ho, Tomaso Poggio, “Face recognition with support vector machines: Global versus Component-based approach”, Massachusetts Institute of Technology, Center for biological and Computational learning.

Classification with Active Learning Method in Relevance Feedback for Content-Based Image Retrieval

Suresh Thommandru
Asst. Professor, Dept. of IT,
Lakireddy Bali Reddy College of Engineering
Mylavaram, India
tsuresh@lbrce.ac.in

Dr. D. Naga Raju
Professor & HOD, Dept. of IT,
Lakireddy Bali Reddy College of Engineering
Mylavaram, India
hodit@lbrce.ac.in

Abstract—The proposed strategy, enhance the attainment of extracting images of content based from the large set of images. To drive the applicability assessment for extracting the images from the large number of images in the scheme of support vector machine (SVM) classifier. The method gives the group of relevant and irrelevant images by processing the ambivalence, assortment of images in the archive. Ambivalence and assortment criteria's target is the selecting representative images in the archive. The proposed method, process the ambivalence and assortment based on 2 steps. The support vector machine used to retrieve the ambivalence images. For retrieving assortment images from the set of ambivalent images using Marginal sampling (MS) method. Proposed method diminishes the difficulties to retrieve the relevant and irrelevant images.

Keywords—*support vector machine; margin sampling; remote sensing images; Active Learning; Relevance Feedback.*

I. INTRODUCTION

The satellite images consisting of single data and spurious data with a series of modifications, those are a large number of images. The confrontation in remote sensing is the decisive and rigorous retrieval of remote sensing images from the archive which are require to user needs. Related works are used keywords or tags like time of acquisition, topographical locations logged in the archive. An impediment in the attainment of virtue, of mostly depend on the availability and aspect of tags. It is very extravagant to get the ambivalence to retrieve the images. New researches have obtain the content of images are more relevant and irrelevant

Image retrieval draws elevating attention in remote sensing for efficient approaches to remote sensing image management. Content based image retrieval consists of feature extraction and retrieval modules. Feature extraction infers a good group of countenance for outline and construes images. Retrieval method finds and extracts the images which are identical to the input image. In remote sensing literature, several primitive countenances are grayscale, color, shape, texture and local invariant [12]. Confront is a semantic gap which is a primitive countenance from an image have a very finite effective in representing and evaluating the high-level concept transmit by remote sensing images. This leads to poor Content based image retrieval performance. It will overcome by the proposed method. It has been constructed to iteratively take considerations of user feedback for improving the content based image retrieval. It is applied to retrieve the relevant and irrelevant images to the input image that are acceptable and pessimistic evaluated the samples.

The suggested method (Active Learning (AL) and Relevance Feedback (RF)) can be handled as a binary classification issue which consists of two years. *Class 1:* for relevance images and *class 2:* for irrelevant images. It can be overcome by using support vector machine (SVM) classification in the context of content based image retrieval by training the SVM with existing annotated images of two classes. Relevance Feedback (RF) is finding the images by user involvement. With precise repetitions, number of times by convalescent the SVM model with the contemporary annotated images. Annotating images as relevant and irrelevant is stagnant and expensive. The conventional relevance feedback method is non constructive and efficient in real applications when large size of archives of images.

Active Learning (AL) to scale down the annotation endeavor in relevance feedback by searching the most oxidative images in the excerpts that, when labeled and consists in grouping of relevant and irrelevant images. It will improve retrieval performance. It acts as a finite number of feedbacks repeatedly to hone the content based image retrieval and a scale down elucidation duration. The place of the training set with a finite act of highly informative images. The proposed method has been elaborated in a framework of classification problem for required image. Untagged samples are eminently ambivalence and assortment to each other frequently preferred as descriptive models to be labeled and include in the training group for the categorization of images.

Ambivalence samples are linked with the certainty of the SVM in correctly classifying it. Assortment samples are associated to correlation in the aspect regions i.e distant to each other. Content based image retrieval issues are more critical than the basic classification issues due to *1)* class of irrelevant images (classifier chosen based on dynamic input image) are larger than class of relevant images. Due to most of the images are irrelevant to the input image. *2)* More incomplete number of annotated images are considered to train by the classifier. Due to truancy of many irrelevant images in the training set. *3)* As real image archives with large numbers of images. It results horizon in ambivalence and assortment classes. It leads to unreliable modeling of the issue. Limitations of Active Learning (AL) only assess ambivalence and assortment samples are inefficient for Content based image retrieval issue.

The remaining section of the theme as follows. Section II Related work. Section III Proposed Active Learning Method. Section IV designates the considered data set and composition of content based image retrieval system. Section V describes experimental results and Section VI conclusion.

II. Related Work

Active Learning (AL) in context of support vector machine is the ambivalence and assortment has been processed in two steps. *Step 1:* The most ambivalence images are chosen from the archive. By Marginal Sampling (MS), unlabeled images are selected. In which, images are closest to the current separating plane. *Step 2:* The assortment images are chosen from the set of ambivalent images by considering the distance between each other. Limitations are unable to represent the images when the large amount of images. Content based image retrieval has been crucial when a small number of annotated images for the large quantity of image feature regions. It does involve the retrieval efficiency than the images retrieved through small quantity parts.

It includes Active Learning (AL) to drive relevance feedback in content based image retrieval choice of most descriptive and definitive untagged images have to be labeled. The proposed method evaluates ambivalence and an assortment of images from the archive. To determine the proposed method exploits 2 stages defined in the scheme of support vector machine. *Stage 1:* By using Marginal sampling (MS), ambivalent images are selected. *Stage 2:* The assortment images among the ambivalence images are selected from the high density regions is by using clustering strategy. It assesses the variety of the untagged images from the aspect regions to commute the excerpt of images to be marked.

New things in proposed Active Learning method for Relevance Feedback in content based image retrieval are *1)* New approaches are used to extract the most informative and representative images in the ambience of CBIR. *2)* To evaluate the representativeness of the images and to tagging the images by a prior term of distribution considering the large amount of unlabeled images in aspect regions. The advantage of using this effectively overcomes the issues caused by inadequate numbers of labels samples in relevance feedback. These are appropriate or effective for remote sensing image retrieval.

Histogram Intersection (HI) Kernel is image retrieving process for proximity metric of image aspects in kernel region. It has not been processed yet. These attempts to carry through an archive of geographical images and demonstrated the efficiency of the proposed method.

III. Proposed Method

Assume by considering an archive "β" made up of very huge number of "I" images.

$$\{X_1, X_2, \dots\dots, X_R\} \;\; where\ X_i\ is\ the\ i^{th}\ image$$

Defined as $\{X_i^1, X_i^2, \dots\dots, X_i^L,\}$ $\quad i = 1, 2, \dots, R$
Aspects representing the content of $i^{th}\ image$

$$X_i^n, \quad n = 1, 2, \dots\dots, N$$

N is the total number of aspects.

Assume $X_q = \{X_q^1, X_q^2, \dots\dots, X_q^L,\}$ be input image

User selected from the archive β (i.e. $X_q \in \beta$) or not existed in archive β (i.e. $X_q \notin \beta$).

Proposed method consists of 3 levels. 1) Finite Level of aspects (features) are extracted from both input image and images in the archive. 2) Defining the Training set module which consists the initial finite Training set of relevant and irrelevant images with regard to the input. 3) Final module returns the δ of images and refines the training set T characterizes by preceding module.

Target is *i)* Relevance feedback (RF) driven by Active Learning (AL) module due to vital parts for assessment of content based image retrieval. *ii)* Feature Extraction. *iii)* Options take into consideration for measuring the proximity of image aspect regions in the proposed system. Active Learning is repeatedly enhanced the size of labeled training set T, excerpts the informative images from the archive β to annotate them.

For relevance feedback (RF) repeatedly the informative untagged images are submitted to classifier are chosen based on the Active Learning method, labeled by support vector machine and append to the present training set T, The support vector machine retained with images and moved from β to **T**. Original training set T should consists of some labeled images for train the classifier and it will append repeatedly the

instructive images which are selected from β. For each repetition the instructive image is confirm whether it is proper classified or not. The user will involve into this process of confirm the instructive image and it will repeated until user satisfaction. This process cost might be reduced by avoiding the replicated images. The efficient class models predicted on quality training set based classification rules can be attained precise retrieval accuracy.

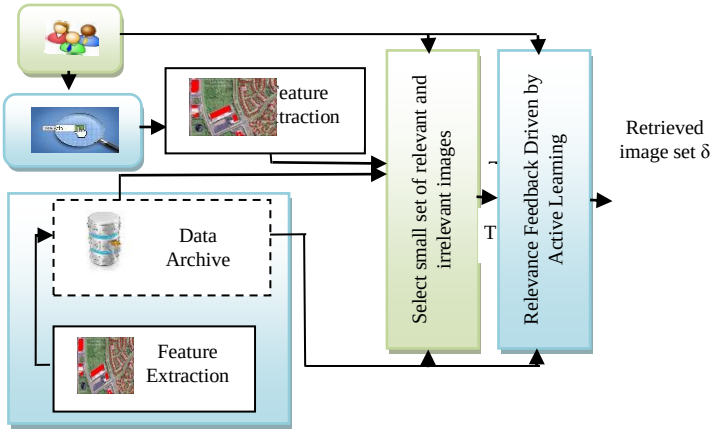


Fig. 1. Content based Image retrieval(CBIR) with Relevance Feedback(RF) driven by Active Learning(AL)

The efficiency of proposed method depends on the potential of Active Learning (AL) method treated to select instructive and representativeness images to be labeled. The user has to involve reaching the final result. To enrich the training set in every repetition of Relevance Feedback in content based image retrieval. Within less number of relevance feedback repetitions to achieve training set with images labeled as relevant and irrelevant by considering query image. And retrieve images with high accuracy.

Active Learning

Input: Set of Positive Target samples Q, Set of Negative Samples R

Set of Un-annotated Data S

Repeat

$L = Q + R$

$x = \arg\min_{x \in S} \left(1 - \frac{P(x|+)}{P(x)}\right) \text{sgn}\left(\frac{1}{2} - \frac{P(x|+)}{P(x)}\right)$

Get label of x from user

If x is labeled as target by user then

$Q \leftarrow Q \cup \{s\}$

else

$R \leftarrow R \cup \{s\}$

end if

Perform Learning using the new training set $L = Q + R$

Until target sample is met.

Select a set $S = \{X_1, X_2, \ldots, X_n\}$ of n images for each relevance feedback repetition as *1)* ambivalent assessing according to Margin sampling (MS). *2)* Assortment might possible for each

Compute minimal distance over the C classes. The result is

images. *3)* Placed in the large portions of the image aspects regions. *2)* and *3)* are evaluated by clustering strategy.

Proposed Active Learning method evaluates the process by strategy based on below steps to choose the set S of images. *Step 1:* $m > n$ ambivalent images are selected according to margin sampling strategy from archive β. *Step 2:* $(m > n > 1)$ assorted images n are among these m ambivalent images are selected from large number of images.

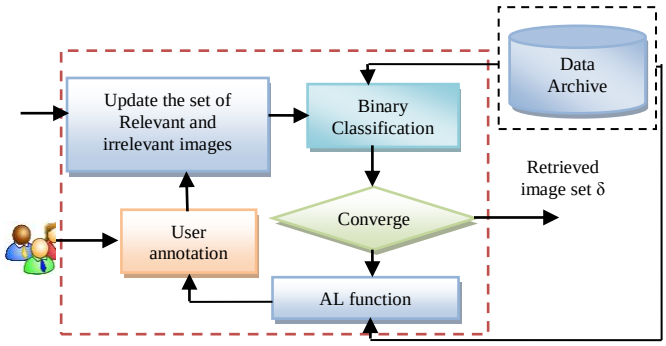


Fig. 2. Process of Active Learning

Step 1: Support vector machine binary classification properties are used for selecting unlabelled images which have maximum ambivalent on their accurate target class. Images which have less probability, they have to be classified accurately by the classifier. These are included in the training set for separating relevant and irrelevant images closed to the separating plane which have low certainty to be accurately classified by Margin Sampling (MS) method. It selects the unlabelled samples close to the separating hyperplane have lowest certainty.

The binary classification support vector machine is trained by using preceding set of relevant and irrelevant images. And then the utilitarian distance between the unlabelled images and the separating hyperplane are predicted.

Margin sampling by Support vector (SV)

Input: Initial training set T, Set of candidates Q, Number of classes C, Pixels to add at every iteration P.

Repeat

Train current classifier with current training set T.

Compute test error of the current classifier.

Repeat

Repeat

Compute the distance to the margin for the candidate Q_i for class C using $f(Q_i) = \text{sign}\left(\sum_{j=1}^{n} y_j \alpha_j\right)$. The result is a $(m \times C)$ distance matrix.

Where m are class membership candidates.

Select the support vector j that minimizes $A(x_j, Q_j)$. The result is a $(m \times C)$ support vector list(SV).

Until each candidate $Q_i$to append

Until each class C

a $(m \times C)$ distance vector.

Append Q_i to a provisional list L.
Repeat
if $SV_i \neq SV_{i-1}$ **then**
Append the candidate related to $\min_{Q \in L} |f(Q_i)|$ to the best candidate list B.
Refresh L.
Append Q_i to a provisional list L.
Else
Append Q_i to a provisional list L
end if
Until i reached to m
Label the P pixels associated with minimal distance in B.
Update T with the P chosen pixels. Remove the selected pixels from Q.
Until end of the list.

Set of "m" images $S^{Uncertain} = \{X_1, X_2, \dots, X_m\}$ where $X_i = \{x_i^1, x_i^2, \dots, x_i^L\}$ is ithuncertain image close to separating hyperplane is seleted. L is primitive features. Choosing the "m" value is significant for determine the efficiency of proposed AL method, when "m" is high value, excerpts images with low level of certainty. "m" is small value, neglecting highly uncertain images. "m = 4h" based on AL literature.

Step 2: To select "h" images from the set $S^{\text{Uncertain}}$ of maximum ambivalent images those are assortment to each other from the large number of sample images. Excepting ambivalent images from the large number of images, cause by more proximity targets having in the images feature regions. The Efficient strategy will optimize on the whole retrieval error. This is an efficient way to assess assortment in the large number of images. Due to unlabelled ambivalent images from different regions are contained sparse in aspect space. Those are treated as assortment images.

Visual words container is take advantage of the information of limited aspects extracted by transformation of interest point in the image. Transformation of image points is translation, rotation and scaling the objects. It is very efficient and reliable in image retrieval. For these transformations descriptors are used to represent local interest points in the image and descriptors are describing the regions of the image. Histogram Intersect (HI) kernel used To evaluate the proximities of visual words container representations of the images in archive. Also, these properties are obtained from the support vector machine and the proposed method.

To measure the proximities between images X_i and $X_j$as $X_i = \{x_i^1, x_i^2, \dots, x_i^L\}$ and $X_j = \{x_j^1, x_j^2, \dots, x_j^L\}$ the Histogram Intersect (HI) kernel can be evaluated as

$$Q(X_i, X_j) = \sum_{l=1}^{L} min(X_i^l, X_j^l)$$

Here $X_i^n \in X_i$, $n = 1, 2, \dots, N$ and $X_j^n \in X_j$, $n = 1, 2, \dots, N$ histogram aspects.

It is a positive definite parameter-free kernel for non-negative feature.

IV. Experiment Results

Data set consist of 9961 images of various categories of vehicles. Samples taken in to 6 X 6 matrix form and validated the retrieved images based on precision.

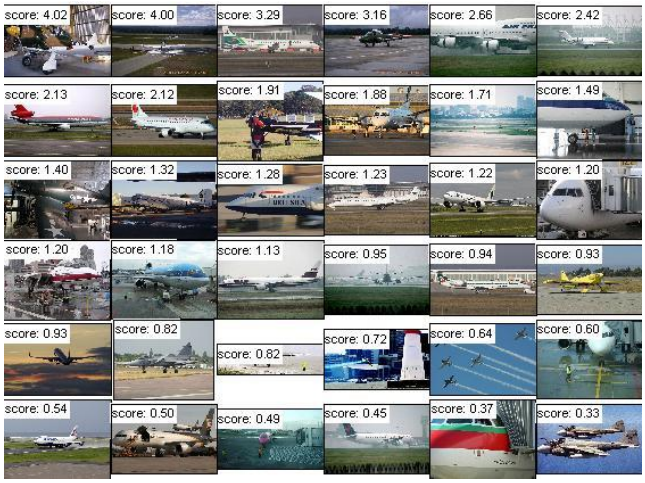


Fig. 3. Retrieved and ranked 36 images as 6 x 6 matrix form.

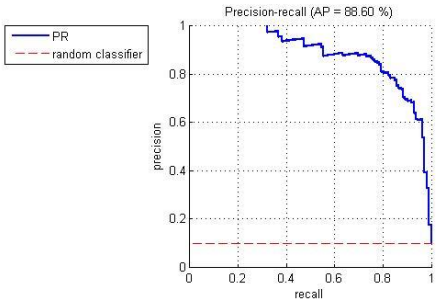


Fig. 4. Precision-recal for Proposed and random sampling.

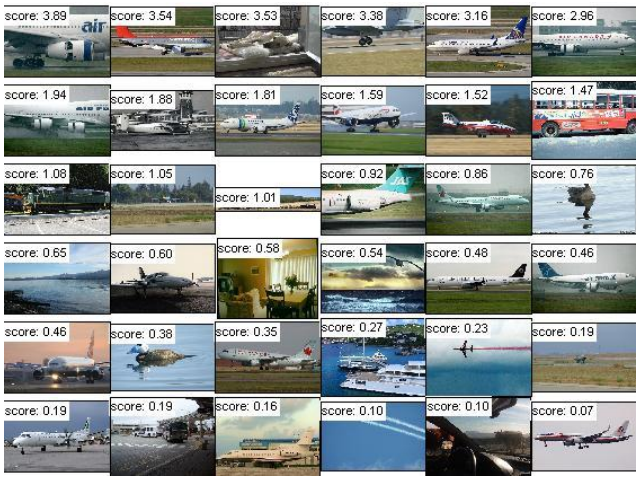


Fig. 5. Retrieved and ranked 36 image for airplane as 6 x 6 matrix form.

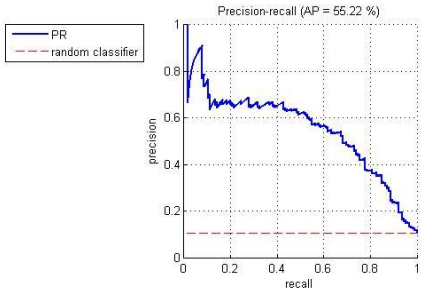


Fig. 6. Precision-recal for Proposed and random sampling.

V. CONCLUSION & FUTURE WORK

AL approach assess collectively standards primarily based on 2 successive steps: most uncertainty images are excerpted from the large set of images by using support vector machine method. Images of each remote areas of picture function area in archive are selected from the maximum uncertainty photographs by way of margin sampling method. Active Learning (AL) technique is to derive the Relevance Feedback (RF) devote to decrease problem of irregular and tendentious set of applicable images.

Dataset of USGS National remote sensing data from http://vision.ucmerced.edu/data/sets/landuse.html. This consists of 2100 images with 21 categories to retrieval and validating the images.

REFERENCES

[1] D.-H. Kim and C.-W. Chung, "Qcluster: Relevence Feedback Using Adaptive Clustering for Content Based Retrieval," *Proc. ACM Conference on Management of Data, 2003.*

.

[2] S.-F. Chang, J.R. Smith, M. Beigi, and A. Benitez, "Visual Information Retrieval from Large Distributed Online Repositories," Comm. ACM, vol. 40, no. 12, pp. 63-71, 1997.

[3] G. Ciocca and R. Schettini, ªUsing a Relevance Feedback Mechanism to Improve Content-Based Image Retrieval,º Proc. Visual '99: Information and Information Systems, pp. 107-114, 1999.

[4] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," IEEE Trans. Geosci. Remote Sens., vol. 51, no.2, pp. 818-832.

[5] Y. Rui, T. S. Huang, S. Mehrotra and M. Ortega, "A relevance feedback architecture in content-based multimedia information retrieval systems", *Proc of IEEE Workshop on Content-based Access of Image ana Video Libraries in conjunction with IEEE CVPR '97.*

[6] X. Huang and L. Zhang, "An SVM ensemble approach combining spectral, structural, and semantic features for the classification of highresolution remotely sensed imagery," IEEE Trans. Geosci. Remote Sens., vol. 51, no. 1, pp. 257–272, Jan. 2013.

[7] G. M. Foody, A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd, "Training set size requirements for the classification of a specific class," Remote Sens. Environ., vol. 104, no. 1, pp. 1–14, 2006.

[8] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classi- fication," IEEE J. Sel. Topics Signal Process., vol. 5, no. 3, pp. 606–617, Mar. 2011

[9] G. Druck, B. Settles, and A. McCallum, "Active learning by labeling features," in Proc. Conference on Empirical Methods in Natural Language Processing, Singapore, 2009, pp. 81–90.

[10] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," Journal of Machine Learning Research, vol. 2, pp. 45–66, March 2002.

[11] L.Bruzzone, C. Persello, and B. Demir, "Active Learning methods in clssification of remote sensing images, "in Signal and Image Processing for Remote Sensing, 2nd ed, C. H. Chen, Ed. Boca Raton, FL, USA: CRC press, 2012, ch. 15, pp. 303-323.

[12] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifier," in Proc. 5th ACM Wkshp. Comput. Learning Theory, Pittsburgh, PA, July 1992, pp. 144–152.

[13] O. Chapelle, S. P. Haffner, and V. Vapnik. "Support vector machines for histogram-based image classification." IEEE Trans. on Neural Networks, 10(5):1055–1064, May 1999.

[14] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. Journal of the American Society of Information Science, 41(4):288–297, 1990.

[15] S. J. Huang, R. Jin, and Z. H. Zhou, "Active learning by querying informative and representative examples," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 892–900.

Protein Secondary Structure Extraction using Bag of Words Model

K. Sushma
PG Scholar,
Department of Computer Science and Engineering
V.R.Siddhartha Engineering College, Andhra Pradesh,India
(email: sushma.kolli01@gmail.com)

Dr.K. Suvarna Vani
Professor,
Department of Computer Science and Engineering
V.R.Siddhartha Engineering College, Andhra Pradesh, India
(email:suvarnavanik@vrsiddhartha.ac.in)

***Abstract*—The protein secondary structure extraction is one of the major areas of interest in bio-informatics, theoretical chemistry, biotechnology and it is highly important in medicine. Protein secondary structure plays an important role in 3D structure prediction. Protein contact map is a 2-dimensional representation of the protein tertiary structure. Protein contact map represents the distance between all possible amino acid residue pairs of a 3D protein structure using a binary 2D matrix. Hence instead of extracting features from the primary amino acid sequence, an algorithm is proposed to extract pattern features from the protein contact maps. The contact map provides most useful information about the protein tertiary structure. Demonstrating the fact that analysis of contact maps can yield better insights for protein structure extraction. The main aim of this work is how the text mining techniques are applied for extracting the protein secondary structure.**

***Index Terms*—Protein, secondary structure extraction, protein contact maps, text mining, Structure Prediction, Bioinformatics, protein tertiary structure.**

I. Introduction

Bioinformatics is a coming forth interdisciplinary field. It conducts with the computational methods and analytic thinking of biological information: genes, genomes, and proteins, cell signaling and metabolic pathways.

Amino acids constitute the proteins. The purpose of a protein is dependent on arranging the amino acids. Proteins are elongate successiveness of amino acids. Amino acids disagree only in side chains and link via peptide bond. Proteins are successiveness of amino acids. Each building block of a protein is called an amino acid remainder because it is the remainder of every amino acid that figures the protein by missing a water molecule. Protein structure [2] in different sizes ranges from tens to several thousand remainders. By physical size, proteins are separated as nanoparticles, between 1100 nm [3]. A protein may have reversible morphological changes in doing its biological function. The optional structures of the same protein are concerned to as unlike conformations, and transitions amongst them are called configurationally changes.

Protein structure is the bimolecular theatrical performance of protein molecule. Proteins are linear polymers constructed from 20 unlike amino acids [2] each of them contributes a common morphological feature. Proteins are similar long molecular chains and are significant component of most biochemical procedures. Protein chains congregation into singular and tightly carried global constructions which are called folds. Each particular episode of amino acids determines the proteins unique congregation. The geometry of a protein congregation finds out its biological function.

Protein structure contains different levels of organization:

1) Primary Structure
2) Secondary Structure
   - Helices
   - Beta Sheets
   - Coils
3) Ternary Structures
4) Quaternary Structures

A. Primary Structure

The primary structure [6] of a protein as shown in the figure 1 refers to the successiveness of amino acids in the polypeptide strand. The primary structure is admitted in concert by covalent bonds such as Peptide bonds.

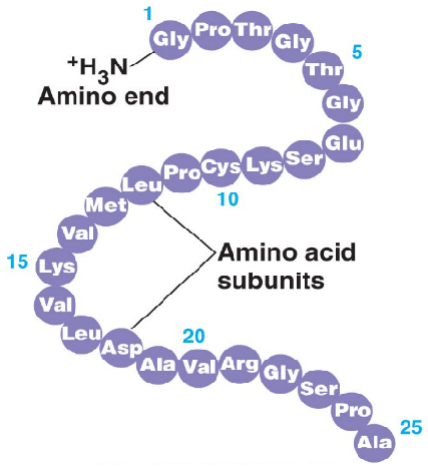


Fig. 1: Primary structure

B. Secondary Structure

Secondary structures [4] concern to the extremely steady local sub-structures. There are two cases of the secondary structures .They are alpha helix and beta sheets. These secondary structures are determined by patterns of hydrogen bonds amongst the main-chain peptide radicals.

Alpha Helix [4] can be organized by creating a rope coil in left hand guidance. In the example of a protein the rope would be constituted by the N-C-C backbone of the polypeptide chain as shown in the figure 2.

Fig. 2: Alpha Helix

Beta Sheets [4] comprise of beta chains associated laterally by at least two or three backbone hydrogen bonds, organizing a generally twisted, pleated sheet. A beta strand (also strand) is stretches of polypeptide strand typically 3 to 10 amino acids long with backbone in an extensive conformation.

Beta Sheets [4] is either in the direction of parallel and anti-parallel ways. Anti-parallel beta sheets are to a greater extent stable than parallel beta sheets because of the hydrogen bonds in the anti-parallel beta sheets are linear as shown in the figure 3 and figure 4.

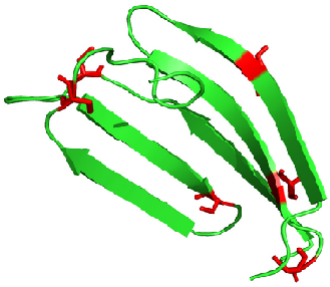


Fig. 3: Parallel and Anti-Parallel beta sheets

Fig. 4: Alpha helix with beta sheets

C. Tertiary Structure

Tertiary structure [5] concerns to the three dimensional representation of a single, double or triple bonded protein structure as shown in the figure 5. The alpha-helixes and beta ruffled-sheets are closed into a compress global structure.

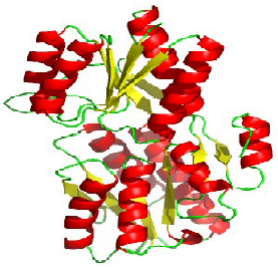


Fig. 5: Tertiary Structure

D. Quaternary Structure

Quaternary structure is the three-dimensional structure of a multi-subunit protein and how the sub-units correspond together. In this circumstance, the quaternary structure is steadied by the same non-covalent interactions and disulfide bonds as the tertiary structure. Composites of two or more polypeptides (i.e. multiple subunits) are called multimers. The data construct requirement to make amino acids is stored in DNA. Amino acid sequence is decided by the DNA sequence. Protein structure is decided by the amino acid sequence. Protein purpose is decided by the protein structure.

A Protein Contact Map [1] constitutes the distance between all potential amino acid residue pairs of a three-dimensional protein construction utilizing a binary two-dimensional matrix. The main advantage with the help of protein contact maps is constant to rotations and translations. Contact maps are used to describe similarity between protein structures.

Text mining also referred to as Text data mining, refers to the process of extracting useful patterns from unstructured textual data. The identification of patterns in text, is based on the Bag of Words (BOW) representation. In Bag-Of-Words [BOW] model the occurrence of (frequency of) each word is used as a feature of training a classifier.

II. Literature

Protein structure prediction is one of the most important goals pursued by bio-informatics and theoretical chemistry, it is also highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes).The prediction of protein secondary structure from amino acid sequence has been attempted since the late 1950s.

The function of a protein depends upon on its structure which itself relies upon on the protein sequence. In this way , understanding how protein sequence folds into a three dimensional structure (3 D-structure) contributes a considerable amount to the understanding of the protein's function, which can help in rational drug design for combating disease and protein engineering.

Prediction of three-dimensional structure of a protein from its amino acid sequences has turned into the most important problems in Bio-informatics community. Unfortunately, the prediction of three-dimensional structure is turned out to be,

NP complete problem [8]. One of the well known approaches in determining the formal problem is to predict the protein secondary structure.

The main objective of secondary structure prediction is to separate irregular structure pattern of residue in amino acid sequences to a class of protein secondary structure elements as a-helix, strand, coil and the remaining types. As of now, there is hundreds thousands known protein sequence. However, only few protein structures have been determined through crystallography and nuclear magnetic resonance image. However, both methods are expensive and practically intricate. Numerous computational approaches prediction problem in the state-of-art.

A neural network was used to predict protein secondary structure. Inputs to the network are encoded by orthogonal encoding schemes that represents amino acids as set of 20-dimensional vectors; they also utilized a window method in which secondary structure is predicted based on its 12 neighbors.

Supporting vector machines and C4.5 decision trees are consolidated to extract important information from the protein structure prediction model, they utilize a multiclass support vector machine proposed by Crammer and Singer. In the first stage of the algorithm, they constructed three discriminant functions. A A new encoding scheme is proposed in which amino acids are encoded based on genetic cordon mapping. They also used dictionary of secondary structure prediction for structure assignment, which diminish the eight states to three states.

Promota and Xiaxia proposed an efficient encoding scheme based on Delaunay, they represented an atom as an fiducial maker. The fiducial maker is then utilized as an input to supporting vector machines. Another algorithm proposed is binding side prediction algorithm that utilized sequence conservation and geometric methods for pocket finding. They compare their algorithm with LIGSITE algorithm and SURNET, and they show cased that, their algorithm finds the best success rate. They also used supporting vector machine to classify which grid points are most likely to bind the ligands base on the properties of grid point. Wang and Juan proposed a new method, which take into account physical chemical properties and structural properties. Vectors are utilized to encode each amino acid. Their encoding takes advantage of hydrophobic interaction; single number encodes every amino acid; the numbers indicate the hydrophobic properties of the 20 amino acids. In the second encoding scheme, they used three-dimensional vector to code each amino acids; the first unit vector represented a-helix, and the second unit vector represented strand and the last unit vector represented coil.

Xiao-Long Ji,Quan Cheng,Qing Wang proposed an progressed supporting vector machine to optimize a Radial basis function kernel (RBF). Their idea is, based on the understanding that, classification in supporting vector is sensitive to the kernel width if the width is expansive, the instance tends to be very similar and likewise, if the width is small, the instance tends to be dissimilar. They scale the width of the Radial basis function (RBF) in a discrete dependent way.

III. Methodology

A. RS126 Data Set

This is one of the benchmark dataset created for evaluating secondary structure prediction schemes by Rost and Sander. This dataset contains 126 proteins which did not share sequence identity more than 25 percentage over a length of at least 80 residues.Later Cuff and Barton showed that the similarity between sequence was more than 25 percentage using more advanced alignment schemes. List of 126 Proteins in RS126 dataset are shown in figure 6.

PDB ID	Chain	PDB ID	Chain	PDB ID	Chain	PDB ID	Chain	PDB ID	Chain
1BMV	1	2UTG	A	4SGB	I	1PYP	-	3CD4	-
4RHV	1	3GAP	A	1MCP	L	1RBP	-	3CLA	-
1BMV	2	3HMG	A	2OR1	L	1RHD	-	3CLN	-
1R09	2	3TIM	A	1GD1	O	1S01	-	3EBX	-
1LMB	3	4SDH	A	2TMV	P	1SH1	-	3ICB	-
4RHV	3	4TS1	A	2WRP	R	1UBQ	-	3PGM	-
2MEV	4	4XIA	A	5CYT	R	2AAT	-	3RNT	-
4RHV	4	5HVP	A	1ACX	-	2ALP	-	4BP2	-
1BBP	A	7CAT	A	1AZU	-	2CAB	-	4CMS	-
1CDT	A	9API	A	1BDS	-	2CYP	-	4CPV	-
1FXI	A	9WGA	A	1CBH	-	2FOX	-	4GR1	-
1GP1	A	1WSY	B	1CC5	-	2FXB	-	4PFK	-
1IL8	A	2LTN	B	1CRN	-	2GBP	-	4RXN	-
1OVO	A	2SOD	B	1ECA	-	2GCR	-	5LDH	-
1TNF	A	3HMG	B	1ETU	-	2GN5	-	5LYZ	-
1WSY	A	9API	B	1FDX	-	2I1B	-	6ACN	-
256B	A	9INS	B	1FKF	-	2LHB	-	6CPA	-
2AK3	A	1FC2	C	1FND	-	2MHU	-	6CPP	-
2CCY	A	5ER2	E	1GDJ	-	2PCY	-	6CTS	-
2GLS	A	6TMN	E	1HIP	-	2PHH	-	6DFR	-
2HMZ	A	1FDL	H	1L58	-	2SNS	-	6HIR	-
2LTN	A	1CSE	I	1LAP	-	2STV	-	7ICD	-
2PAB	A	1TGS	I	1MRT	-	3AIT	-	7RSA	-
2RSP	A	2TGP	I	1PAZ	-	3B5C	-	8ABP	-
2TSC	A	4CPA	I	1PPT	-	3BLM	-	8ADH	-
9PAP	-								

Fig. 6: list of 126 proteins in RS126 dataset used in Protein Secondary Structure Extraction

A novel approach for identifying the protein structure without any machine learning algorithms with low cost and efficient extraction and compression of protein secondary structure results in identifying 3D patterns of data set for -Helix, -Sheets and coils were organised and compressed based on distance matrix, contact map generation, Heuristic and BOW model.

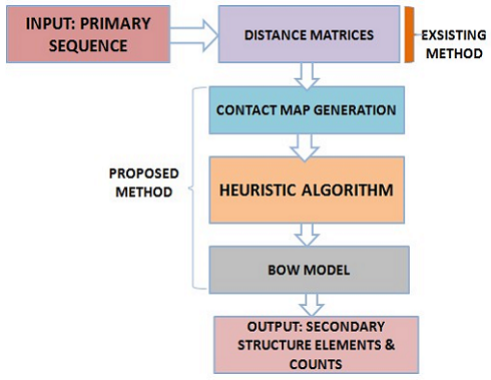


Fig. 7: Block diagram for Protein Secondary Structure Extraction

Primary sequence in fasta format is considered as input. Primary sequence is shown in figure8. The Primary Sequences of Proteins are obtained from Protein Data Bank.

>9INS:B|PDBID|CHAIN|SEQUENCE
FVNQHLCGSHLVEALYLVCGERGFFYTPKA

Fig. 8: Input Sequence

Distill [10] a suite of servers used for the generation of distance matrices. The servers are based on large-scale ensembles of recursive neural networks. By using this Distill server we generate the distance matrices. Distill [10] is used to trace the distances between primary sequence and this helps in contact map generation and the distance matrices were shown in figure 9.

```
0.995658 0.995696 0.977968 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.995696 0.99624 0.995948 0.995396 0.713382 0.688747 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.977968 0.995948 0.996118 0.994687 0.996144 0.980431 0.971122 0 0 0 0 0 0 0 0 0
0 0.995396 0.994687 0.996223 0.995841 0.996216 0.99511 0.988196 0 0 0 0 0 0 0 0 0
0 0.713382 0.996144 0.995841 0.996094 0.99562 0.996121 0.995152 0.988249 0 0 0 0
0 0.688747 0.980431 0.996216 0.99562 0.995947 0.995468 0.996067 0.995245 0.987905
0 0 0.971122 0.99511 0.996121 0.995468 0.995989 0.995253 0.996128 0.995398 0.9892
0 0 0 0.988196 0.995152 0.996067 0.995253 0.995988 0.994854 0.996098 0.995405 0.9
0 0 0 0 0.988249 0.995245 0.996128 0.994854 0.995915 0.995382 0.996088 0.99548 0.
0 0 0 0 0 0.987905 0.995398 0.996098 0.995382 0.996024 0.995396 0.996105 0.995459
0 0 0 0 0 0 0.989279 0.995405 0.996088 0.995396 0.996052 0.995153 0.996107 0.9954
0 0 0 0 0 0 0 0.988766 0.99548 0.996105 0.995153 0.996041 0.995315 0.996125 0.995
0 0 0 0 0 0 0 0 0.989343 0.995459 0.996107 0.995315 0.996071 0.995469 0.996149 0.
0 0 0 0 0 0 0 0 0 0.989625 0.995488 0.996125 0.995469 0.99611 0.995303 0.996141 0
0 0 0 0 0 0 0 0 0 0 0.989138 0.99555 0.996149 0.995303 0.995847 0.995172 0.996087
0 0 0 0 0 0 0 0 0 0 0 0.989149 0.995487 0.996141 0.995172 0.994035 0.993917 0.996
0 0 0 0 0 0 0 0 0 0 0 0 0.99042 0.995594 0.996087 0.993917 0.996022 0.995553 0.99
0 0 0 0 0 0 0 0 0 0 0 0 0 0.989887 0.995312 0.996113 0.995553 0.9955 0.994522 0.9
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.992855 0.995396 0.99614 0.994522 0.992697 0.995106
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.44628 0.993075 0.988745 0.994532 0.995106 0.995839
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.228105 0 0 0.994351 0.994598 0.99613 0.995901 0.9
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.430688 0 0 0.980218 0.995987 0.995901 0.996132 0.
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.874416 0 0 0.979072 0.956455 0.995422 0.995019 0.
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.850606 0.864664 0 0 0.864228 0 0 0.987471 0.995183
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.947942 0.994549 0.996261 0.995651 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.98595 0.995651 0.996243 0.995767
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.99335 0.995767 0.996191 0.99574
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.990669 0.995744 0.996166 0.99
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.992942 0.996055 0.996218 0.
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.564421 0.995437 0.99374 0.9
```

Fig. 9: Distance Matrix

B. Contact Map Generation Algorithm

Protein contact map represents the distance between all possible amino acid residue pairs of a three -Dimensional protein structure using a binary two-Dimensional matrix. A Primary Sequence was considered to generate Protein Secondary Structure for this RS126 dataset was used as a primary data set to extract distance matrices using Distill[10] tool. To generate Contact Map the distance matrices were now tuned to generate contact maps in the form of Sparse matrix. The contact matrix consists of only binary values, hence it is also called as binary matrix.

Algorithm 1 Contact Map Generation Algorithm

Input: Distance Matrices [Distance Matrices are generated from Primary Sequence]
Output: Contact Map
Input parameters: row: row col: column

```
for col ← 0 to n do
    if (val[col]! = 0) AND (val[col] <= 7) then
        matrix[row][col]=1
    else
        matrix[row][col]=0
    end if
end for
row++;
```

C. Heuristic Algorithm

Heuristics involves using an approach of learning and discovery to reach a solution. If you need to develop a solution then it's Heuristic. The term Heuristics[11] is used in algorithms which find solutions among all possible ones. A Heuristic is a short-cut or "Rule of thumb" that gives some guidance on how to do task, but it does not guarantee solutions consistently.

Algorithm 2 Heuristic

Input: Contact Map
Output: Protein Secondary Structure elements
Input parameters: n: total-row-count col:1 $count$:0

```
for i ← 1 to n do
    for j ← col to i + 3 do
        if (matrix[i][j] == 1) AND (count < 4) then
            count=count+1
        else if (count == 4) then
            structure[i]=H
        else if (count == 3) then
            structure[i]=E
        else
            structure[i]=C /* count==2 */
        end if
    end for
    row++
end for
```

D. Bag of Words Model

Bag of Words[12] is also referred as Unordered words with frequencies. Bag is a set that allows multiple occurrences of the same element.Bag of Words model ignores the order of words as well as any punctuation (or) structural information, but retains the number of times each word appears. Bag of Words model is commonly used in methods of document classification, where the frequency of/occurrence of each word is used as a feature of training a classifier. The identification of patterns in text, is based on the Bag of Words representation. Positions and counts(frequency) of Protein

Secondary Structure were identified for Helix, Sheet and coils. These are traced by using BOW model.

Algorithm 3 Bag of Words Algorithm

Input: Protein Secondary Structure Elements [output of heuristic algorithm]

Output: Position and counts of Protein Secondary Structure elements

1: Individual Splitting: The entire data is divided into individual letters.
2: **if** $data[i] == H$ **then**
3: countH=countH+1
4: **else if** $data[i] == E$ **then**
5: countE=countE+1
6: **else**
7: countC=countC+1 /* data[i]==C */
8: **end if**

IV. Results

From generated Distance matrix, Contact map is extracted with a threshold limit of8A , all the above values of limit is tuned to ones and the below are tuned to zeros. The Extracted Contact Map is shown in figure10.

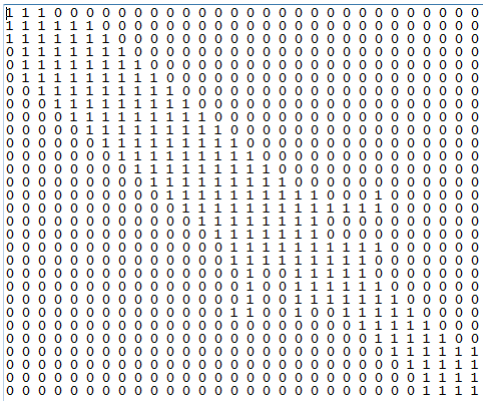


Fig. 10: Contact Map

From generated Contact map all the diagonal values of protein structure are extracted and apply the count based heuristic approach that results in protein structure identification. The result is show cased in Figure11.

Fig. 11: Protein Secondary Structure

From the extracted Protein Secondary Structure elements,Positions and counts of protein structure were identified for Helix, Sheet and coils. These are traced using BOW approach and were optimised or compressed for 3D protein structure generation. The result is show cased in Fuigure12.

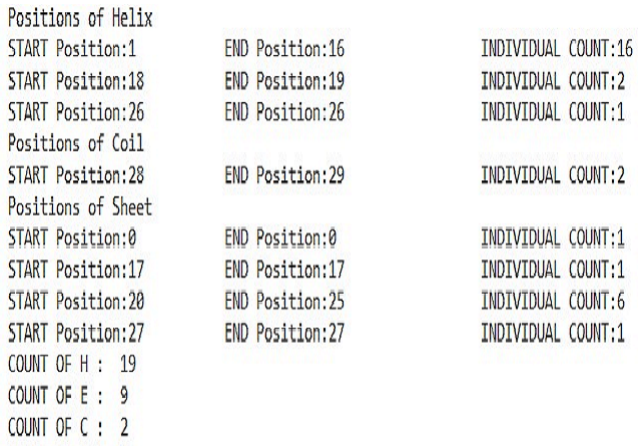


Fig. 12: Protein Secondary Structure element positions and their counts

V. Conclusion

Protein Secondary Structure Extraction is helpful in identifying the Ternary Structure of Protein and was used in bio-informatics for obtaining the information regarding a protein.The protein secondary structure extraction is one of the major areas of interest in bio-informatics, biotechnology and it is highly important in drug discovery and medicine. Here a reference of Contact map generation algorithm, Heuristic algorithm, and Bag of Words model Algorithm combinationally provide a novel algorithm for extracting Protein Secondary Structure. This result is more efficient when compared to classification and other algorithms.

References

[1] M.Vendruscolo, B. Subramanian, I.Kanter, E.Domany, J.Lebowitz, Statistical properties of Contact Maps, PhysRev E 59, 1999.
[2] Pelta, David A., Juan R. Gonzlez, and Marcos MorenoVega."A simple and fast heuristic for protein structure comparison."bioinformatics 9.1, 2008.
[3] Brocchieri L, KarlinS(2005-06-10). "Protein length in Eukaryotic and Prokaryotic proteomes". Nucleic Acid Research 33 (10).
[4] Chiang YS, Gelfand TI, Gelfand IM (2007). "New classification of super secondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage". Proteins. 68(4): 915921.
[5] Barah, Pankaj, and SomdattaSinha. "Analysis of protein folds using protein contact networks." Pramana 71, 2008.
[6] Bock JR1, Gough DA. Predicting protein-protein interactions from protein structure, Bioinformatics, 2001 Oxford University Press.
[7] He J, Harrison R, Tai PC, Pan Y. Rule generation for protein secondary structure prediction with support vector machines and decision tree, IEEE Trans Nanobioscience. 2006 Mar; 5(1):46-53.
[8] Niles A.Pierce and Erik Winfree, Protein Design is NP-hard, Protein Engineering vol.15 pp.779782, 2002.
[9] http://www.ee.unimelb.edu.au/ISSNIP/bioinfo/
[10] http://distillf.ucd.ie/distill/
[11] http://students.ceid.upatras.gr/ papagel/project/
[12] https://en.wikipedia.org.in/wiki/Bag-of-words-model/

Analysis of Breast Cancer Diagnosis using Cytological Images

O. Likhitha
PG Scholar,
Department of Computer Science and Engineering
V.R.Siddhartha Engineering College, Andhra Pradesh,India
(email: likhitha.odugu@gmail.com)

Dr. K. Suvarna Vani
Professor,
Department of Computer Science and Engineering
V.R.Siddhartha Engineering College, Andhra Pradesh, India
(email:suvarnavanik@vrsiddhartha.ac.in)

***Abstract*—Deadly disease that causes many women to breach the walls of death was caused by breast cancer. According to the census 2010-12 due to breast cancer 21.5 percent women were died across the world i..e, more than 100000 women per year. So, to identify the disease an image processing classification process is required to diagnose the patient imaging medical data. To identify and classify the disease of a person in this paper a 6 stage process is proposed. These stages include morphological filtration, gradient conversion, otsuthresholding, fuzzy c means clustering, water shed segmentation and svm classification. All these combinations were used at different levels of process. This process is carried out by using classification of cancer images and identifies the cancer from image medical data with 95.31 percent accurately.**

***Index Terms*—Women, breast cancer, gradient conversion, otsuthresholding, fuzzy c means clustering, water shed segmentation and svm classification.**

I. Introduction

Breast cancer is one of the leading cancer for women world wide. Indications of breast cancer indicates the irregularity , an adjustment fit as a fiddle, skin dimpling, or a red layered patch of skin[1]. In those with removed spread of the ailment, there might be bone torment swollen lymph hubs, shortness of breath, or yellow skin[2]. Hazard components for creating bosom disease include: stoutness, absence of physical activity, drinking liquor, hormone substitution treatment amid menopause, ionizing radiation, early age at first monthly cycle, having youngsters late or not in any manner, more seasoned age, and family history. Around 510 percent of cases are because of qualities acquired from a man's parents,including BRCA1 and BRCA2 among others.Diseases creating from the channels are known as ductal carcinomas, while those creating from lobules are known as lobular carcinomas[1]. What's more, there are more than 18 other subsorts of bosom malignancy. A few tumors create from preobtrusive sores, for example, ductal carcinoma in situ. The determination of bosom malignancy is affirmed by taking a biopsy of the concerning knot. Once the finding is made, further tests are done to figure out whether the malignancy has spread past the bosom and which medications it might react to.The equalization of advantages versus damages of bosom malignancy screening is disputable. A 2013 Cochrane survey expressed that it is misty if mammographic screening benefits increasingly or hurt. A 2009 audit for the US Preventive Services Task Force alternately raloxifene might be utilized as a part of a push to anticipate bosom disease in the individuals who are at high danger of creating it. Surgical expulsion of both bosoms is another safeguard measure in some high hazard ladies. Bosom recreation may happen at the season of surgery or at a later date. In those in whom the growth has spread to different parts of the body, medications are generally gone for enhancing personal satisfaction and solace. Results for bosom growth shift contingent upon the tumor sort, degree of illness, and individual's age. Survival rates in the created world are high, with between 80 percent and 90 percent of those in England and the United States alive for no less than 5 years. In creating nations survival rates are poorer. Around the world, bosom malignancy is the main sort of malignancy in ladies, representing 25 percent of all cases. In 2012 it brought about 1.68 million cases and 522,000 passings. It is more normal in created nations and is more than 100 times more normal in ladies than in men. Around the world, bosom malignancy is the most widely recognized intrusive tumor in ladies. It influences around 12 percent of ladies around the world. (The most normal type of growth is non-obtrusive non-melanoma skin disease; non-obtrusive tumors are for the most part effectively cured, cause not very many passings, and are routinely avoided from tumor measurements.) Breast growth involves 22.9 percent of obtrusive diseases in ladies and 16 percent of every single female tumor. In 2012, it contained 25.2 percent of growths analyzed in ladies, making it the most well-known female growth.

II. Literature

In 2008, breast cancer caused 458,503 deaths worldwide[1]. Lung cancer, the second most common cause of cancer-related death in women, caused 12.8 percent of cancer deaths in women (18.2 percent of all cancer deaths for men and women together).

The incidence of breast cancer varies greatly around the world: it is lowest in less-developed countries and greatest in the more-developed countries[2]. In the twelve world regions, the annual age-standardized incidence rates per 100,000 women are as follows: in Eastern Asia 18, South Central Asia 22, sub-Saharan Africa 22,South-Eastern Asia 26, North Africa and Western Asia 28, South and Central America 42, Eastern Europe 49, Southern Europe 56, Northern Europe 73, Oceania 74, Western Europe 78, and in North America, 90

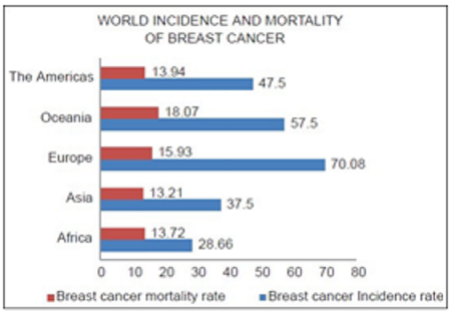


Fig. 1: Rate of breast cancer incidence and mortality worldwide according to 2012 world cancer report

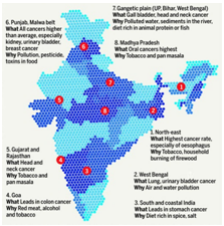


Fig. 2: Geography of cancer in India area wise

Cancer of the female breast was the most common cancer in women being the leading site in Mumbai, Thiruvananthapuram and Dibrugarh[8].

III. Methodology

The purpose of this study is to detect breast cancer from cytological images. Moreover, this system will be merged with the tele medicine. The nuclei detection and water shed segmentation deal with most of the problems related to cell nuclei location detection. Apart from this the cell features are extracted from the data set cytological images.

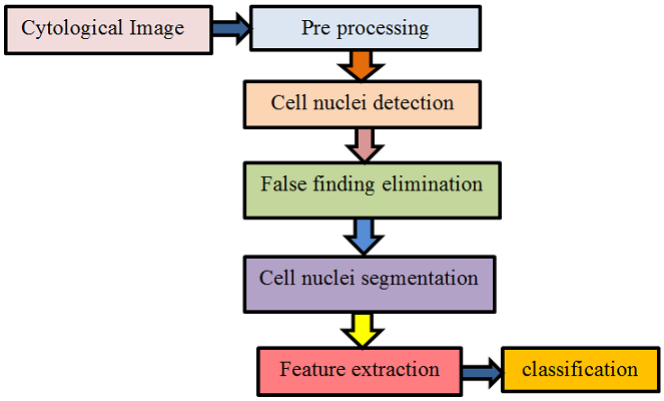


Fig. 3: Block diagram

A. Cytological Image

Cytology refers as a branch of phytology, the medical specialty that deals through the examination of tissue samples with making diagnoses of diseases and conditions from the body. Cytological examinations may be performed on body fluids like blood, urine, and cerebrospinal fluid or on material that is aspirated from the body.

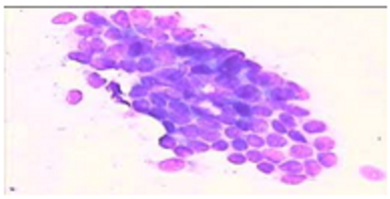


Fig. 4: Example of Cytological image

B. Preprocessing

As the handling time is an exceptionally vital variable in picture preparing, we resized the pictures from a determination of 2560 1920 pixels to 640 480 pixels. At that point, a difference upgrade what's more, edge honing procedure is connected as a lot of pictures has a low difference. In this paper, we utilize basic histogram preparing 0 to be specific, the cumulated sum approach[3] with 1 percent immersion at low and high intensities of the info picture.we apply the contrast-limited adaptive histogram-equalization[4] to enhance the nature of images.After improving the quality of the picture, a dim scale was removed from the shaded picture to be utilized as a part of the following strides of the proposed method.The glow part can be resolved utilizing, i.e.

$$Y = 0.299R + 0.587G + 0.114B. \quad (1)$$

At last, the gradient of the image is evaluated as it will be utilized in the cell nuclei detection and segmentation stages.

C. Cell nuclei detection

Circle Detection: Nuclei locations are detected using circular Hough transform (CHT) and Mexican hat filter. Most of the nuclei will be in elliptical shape but the detection of the elliptical shape nuclei is highly expensive.On the other hand, the shape of the ellipse by a given number of circles can be approximated. The circle detection is simpler in the sense of the required computations because there is only one parameter for circle, which is radius r. The following observations and simplifications form a basis for a nucleus detection algorithm. In this approach, we try to find the circles with different radii in a given feature space. The circular hough transform[6] is used to easily determine the parameters of the circle when a number of points that fall on the perimeter are known.

D. False Finding Elimination

1)Otsus Thresholding Method: We use Otsus thresholding method[7] for removing the markers that belong to noisy circles.We produce a binary mask BW with the regions of interest in the image by thresholding the gray-scale image prepared in the preprocessing stage. We compute a threshold using otsu method that can be used to convert an intensity image to the binary image that selects threshold to reduce

intra class variance of black and white pixels. For detected marker which get the corresponding point in the binary mask and remove that marker when point belongs to corresponding area.

Algorithm 1 Otsu algorithm

Input: Cytological Image
Output: Image detecting circles

1: Calculate histogram for the image
2: Set probability and mean values initially
3: **for** $t \leftarrow 1$ to N **do**
4: update probability and mean
5: **end for**
6: Compute noise estimation
7: Set maximum estimation as threshold limit

2) Fuzzy C-Means Clustering: Fuzzy c-means (FCM) is a method of clustering, that allows one piece of data to belong to two or more cluster. The idea of FCM is to minimize the total weighted mean-square error using the minimum weights: The Fuzzy c-means allows every feature vector that belong to each cluster with a fuzzy truth value (between 0 and 1)The FCM algorithm uses reciprocal distance to compute the fuzzy weights. When a feature vector from two cluster centers is of equal distance, it weights the same on the two clusters[8]. It cannot differentiate the two clusters with different distributions of feature vectors.The Fuzzy C-Means algorithm lumps the two clusters with natural shapes in to large clusters but close boundaries. For some difficult data like WBCD data, it is hard for the Fuzzy C-Means to cluster the very closed classes without taking the help of other mechanisms like removal of small clusters. The Fuzzy c-means uses Gaussian weights that are representative and immune to the outliers. Gaussian weights reflect the distribution of the clusters in the feature vectors. For a feature vector from two prototypes with equal distance, it weighs heavy on the widely distributed cluster than on the narrow distributed cluster.

Algorithm 2 Fuzzy c-means algorithm

Input: Cytological Image
Output: Image detecting cancer cells

1: Initially select cluster centers
2: Initialize U=[uij] matrix, U(0)
3: calculate the center vectors
4: Update U(k) , U(k+1)
5: **if** $(||U(k+1) - U(k)|| < e)$ **then**
6: STOP
7: **end if**
8: otherwise return to step 2.

E. Cell nuclei Segmentation

To separate attached cancer cells in to individual objects watershed transformation[5] is used. Marker-Controlled Watershed Transform: The watershed transform[9] and [10] is defined as a region based segmentation method. As any gray scale image can be considered as topographic surface which regard as the intensity of a pixel as altitude point.The application of the watershed transform in the default form results in the over segmentation of the image because of the presence of artifact and noise.

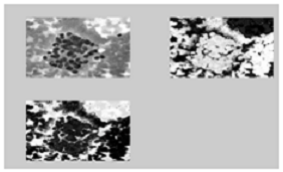


Fig. 5: fuzzy clustered images

F. Feature Extraction

The efficient classification from the total segmented regions of nuclei cells requires the generation of features of good discriminative capacity. The nuclei areas founded of the nuclei enclosed by the detected boundaries, features having the detected regions of the shape and texture can be determined easily. In this work we use ten shape based features and two textural features[11]. The values obtained for the above features yield a well differentiation between the healthy cells and cancerous cells. These features are proposed as input data for the classification phase. Shape features: The boundaries detected for the nuclei are expected to present an ellipse-like shape and several features to describe this characters have chosen. Ten features are calculated from the extracted shape of region boundary, such as perimeter, compactness, smoothness, eccentricity, solidity, equivalent diameter, extent, major axis length, and minor axis length.All these features are useful to classify.The values obtained for the shape features yield a good differentiation between the healthy and cancerous cells. These features are taken as input for the classification phase. Textural features: From the texture of the cell nucleus two features are calculated i.e the standard deviation in both the gray scale of the RGB color model and YCbCr color model of the gray scale that is Y-level

G. Classification

Classification is task of providing an individual item in to certain category which is called as a class. The task of the classification is done by a classifier which takes input as a feature vector and gives the category as the output to which the object belongs. The feature vector is set of features extracted through the input data.In this paper the feature vector represents 12 features extracted for each nucleus as illustrated above in the feature extraction phase. The classification step was done using well known supervised classification technique called Support Vector Machine(SVM)[12].

IV. Results

The cytological image of the breast cancer is taken as the input as shown in fig 6

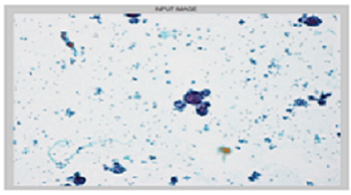


Fig. 6: Input image

Figure 7 shows the Histogram streching which is performed to improve image quality.

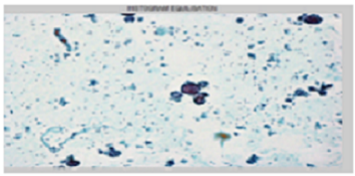


Fig. 7: Histogram Streching

Figure 8 shows the gray scale extraction which is used as input in next steps.

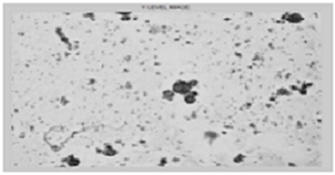


Fig. 8: y-level

Figure 9 shows the gradient image which is taken as input to cell nuclei detection and segmentation stages.

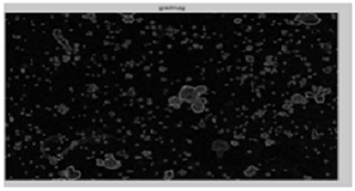


Fig. 9: Gradient image

Circular hough transform is used to detect the nuclei locations.Nuclei locations are detected as shown in fig 10

Otsu thresholding is performed to eliminate the noisy circles.

FCM is used to detect the nuclei markers where red markers indicate cancer cells and blue markers indicate blood cells as shown in Figure 13

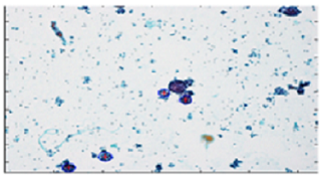


Fig. 10: Circular Hough Transform

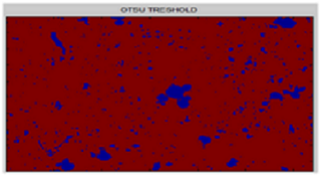


Fig. 11: Otsu Thresholding

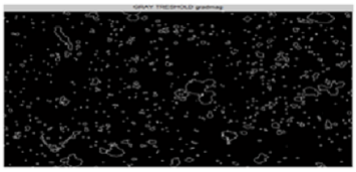


Fig. 12: Otsu thresholding

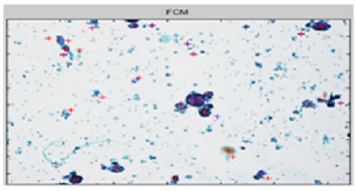


Fig. 13: Fuzzy c-means

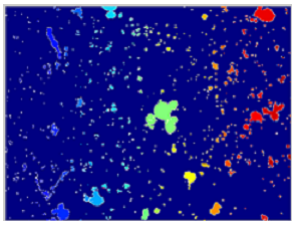


Fig. 14: WATERSHED

Watershed transform is performed to divide attached cancer cells in to individual objects

SVM classification is chosen to achieve high accuracy and identifies the cancer from image medical data with 95.31 accuracy.

V. Conclusion

In this work we have developed automated system for detection of breast cancer from the given cytological images. Through this work the detection of nuclei markers problem has been eliminated. The performed experiments shows the Hough transfrom used for circle detection that can be

effectively used for the pre segmentation of cell nuclei in cytological images. Otsu threshold eliminates all of the noisy circles, whereas with out any loss of true nuclei markers. In addition, the FCM algorithm produces high accuracy for the clustering of nuclei markers corresponds to blood cells and true nuclei. The proposed work shows the outcome of the watershed transform is accurate nuclei boundaries. The use of the proposed detection and segmentation method is that it is fully automated and it is suitable for images with a high degree of noise and blood cells and cell overlapping, as it detects successfully not only nuclei cells but also the nuclei in cell clusters. For the feature extraction ten shape based features and two textural features are used. For the classification phase SVM classification is used to achieve high accuracy and identifies the cancer with 95.4 percent accuracy.

REFERENCES

[1] Breast cancer treatment NCI,23 May,2014
[2] Saunders, Christobel, Jassal, Breast Cancer Oxford Universty press.
[3] Gonzalez R.C and R. E Woods, Digital Image Processing, 2nd ed. Boston, MA, USA: Addison Wesley, 2001.
[4] Zuiderveld .K, Contrast limited adaptive histogram equalization, in Graphics Gems IV. San Diego, CA,year 1994.
[5] Beucher.S and C. Lantuejoul, Use of Watersheds in Contour Detection, Sep 1979
[6] M. Roushdy, Detecting coins with different radii based on Hough transform in noisy and deformed image, J. Graphics, Vision Image Process Apr. 2007.
[7] Otsu. N, A threshold selection method from gray-level histograms, Jan. 1979.
[8] M. E. Plissiti, and A. Charchanti, Automated detection of cell nuclei in pap smear images using the morphological reconstruction and clustering, 241, Mar. 2011
[9] M. Wang and X. Zhou, Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy, Bioinformatics,2008.
[10] M. E. Plissiti, H. Nikou, and A. Charchanti, Watershed-based segmentation of nuclei cell boundaries in Pap smear images, in Proc. IEEE Int. Conf. ITAB, 2010, pp. 14.
[11] W. H. Wolberg, W. N. Street, Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates, Staging Cancer, Mar. 1994.
[12] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Method, 2000.

Analysis of Different Bioinformatics Analytic Procedures in Biomedical Big Data Evaluation

P. Udayaraju[1] , Dr. K. Suvarnavani[2] , Dr. Chandra Sekhar Vasamsetty[3]
[1]Assistant Professor Department of CSE, SRKR Engineering College, Bhimavaram, AP, India
[2]Professor, Department of CSE, VR Siddhartha Engineering College, Kanuru, Vijayawada, AP, India
3Associate Professor Department of CSE, SRKR Engineering College, Bhimavaram, AP, India
Email: [1] udayaraju8910@gmail.com , [2] suvarnavanik@gmail.com, [3] chandu.vasamsetty@gmail.com .

Abstract: Large amount of data produced from healthcare informatics and bioinformatics has been grown to be quite vast in analysis of big data based on knowledge grained with possibilities arranged in real time data evaluation. Because of increasing trend towards personalized and precision medicine biomedical data from various sources in different structural dimensions. Healthcare and bioinformatics provides clear disciplinary intent to combine data & knowledge with available information based on effective decision making in clinics and translational research. To defect on different representations related to role of data analysis in healthcare and biomedical informatics. In this paper we analyze different approaches for big data analysis with respect to biomedical and healthcare informatics data collected at multiple levels data processing. Furthermore gathering data from different levels, different levels queries addressed in human scale biology, clinical scale and epidemic data representation. We review recent works and break thoughts of big data applications processing in healthcare domains and summarize the challenges to improve big data application development in bioinformatics and health care informatics.

Index Terms: Big Data, Data driven application, Health informatics, Bioinformatics, State-of-the Art, Public health informatics, Translational Bio Informatics.

I. INTRODCUTION

Health informatics has changed and started to solve and handle progressive knowledge of big data analysis with preferable presentation. By performing data mining with big data analytics diagnosing helping all the patients in both health informatics and bioinformatics. The field of bioinformatics, health informatics are the cusp to support period of date and entering new era as a technology to solve and handle Bid Data about unlimited potential for information increase in real time application development [1][2]. Big data analytics are helping to realize the diagnosing, treating and healing with need of healthcare informatics and bioinformatics.

Health Bioinformatics is a combination of data science and software engineering inside the domain of medicinal services. There are various ebb and flow zones of examination inside the field of Health Informatics, including Bioinformatics, Image Informatics (e.g. Neuroinformatics), Clinical Informatics, Public Health Informatics, furthermore Translational BioInformatics (TBI) [4]. Research done in Health Informatics (as in all its subfields) can range from information obtaining, recovery, stockpiling, and investigation utilizing information mining procedures, and so on. Nonetheless, the extent of this study will be examination that utilization information mining with a specific end goal to answer questions all through the different levels of health. Various research methods done on health informatics uses information from some required point of levels in human existence, Bioinformatics use molecular level of data, neuro informatics uses semantic level of data, clinical informatics uses patient level of data and lastly public level informatics uses population data in real time application development with processing of data management. In this study various sub-ordinates were progressed for health and bioinformatics are : " Big data evaluation in health informatics", which represents overall description of health informatics, "Levels of health informatics", which discuss various sub environments in health informatics, "Use Micro level molecular data", public health utilization processed population data [3][4].

However, levels of data suspended research studies in individual biomedical questions of study attempts to answer where each question associated with scope data level presented in development of data levels. The main tissue data level is analogous scope to human biology scale queries, the scope of patient data is related to biomedical with clinical queries. Healthcare is an important to economy for society to its emotional and dream able to vision of sustainable improvement in both physical mental health of its individual service orientation. Therefore numbers of techniques were improved to handle, analyze healthcare system in favorable environment. Large volumes of data from bioinformatics and health informatics coupled with emerging analytics are estimated to implement future preventive, predictive and personalized health informatics in real time data sharing [6]. Bioinformatics provides to different research authors to store data such as DNA sequence with analysis and interpretation for excellent analysis and interpretation on forms of databases. Bioinformatics has enabled scientists by professional researchers, in this paper bioinformatics provides analysis of Gene Expression Data, DNA and Protein sequences, protein-to-protein interaction by molecular analysis and Gene Ontology Hierarchy in both health informatics and bioinformatics with sequential development [5][8].

Remaining of this paper organized as follows: Section 2 describes general implementation of literature of bioinformatics with health informatics implementation. Sections 3 formalize to develop Visual Analytical Approach to handle Gene Expression Data with implementation procedure. Secure 4 define evolutionary analysis of DNA protein interaction sequences in heath data. Section 5 predicting protein functions in protein-to-protein interaction in biomedical data. Section 6 defects web based implementation for interesting gene interaction using Gene Ontology hierarchy. Section 7 concludes overall analysis of bioinformatics with above considerations.

II. LITERATURE SURVEY

In this section big data refers to tools and implementation and procedures with organizational to create manipulate very large data sets and storage facilities. Literature of big data analysis is as follows:

Demchenko et al.[1] depicts huge data by five versus: Quantity, Speed, Wide range, Veracity and Value. Amount shows the extensive measures of data utilized. Speed shows the rate at which new data is created. Wide range demonstrates the level of the multifaceted nature of data. Veracity is utilized to take a gander at the unwavering quality of the data. Butte et al. [2] analyzed that few TBI focuses on outlined in JAMIA which conmibe natural details with therapeutic information to achieve regenerative improves as more details points are tried. Makers comment that TBI started from an discovery done by a little collecting who found how to go over any hurdle between computational technology and solution.

Sarkar et al. investigates that there are three areas of important quest for TBI: determining the nuclear level(genotype) sways on growth and development of disease, understanding common reliability between sub-atomic, phenotype and environmental connections crosswise over various population, taking in the effect of helpful systems as can be calculated by sub-atomic biomarkers [7][9][10]. They believe in that TBI is an important position to perhaps decide a significant section of the questions of complicated health problems or any of the other evaluation with the explosion of both nuclear stage details and biomedical details.

Numerous issues on Big Information projects can be settled by e-Science which requires network preparing. e-Sciences incorporate compound science, bio-informatics, earth sciences and open models. It likewise gives advances which permit assigned participation, for example, the Access Lines. Molecule science has an all around created e-Science foundation specifically in view of its requirement for adequate preparing highlights for contextual investigation of results and capacity of information through the European Organization for Nuclear Research (CERN) Huge Hadron Collider, which began taking information amid 2009. E-Science is a major thought with numerous sub-fields, for example, e-Social Technology which can be viewed as a higher improvement in e-Science. It plays out a

section as a piece of open science to assemble, handle, and investigate the general population and behavioral information. Other Big Information programs relies on upon numerous therapeutic callings like stargazing, ecological science, solution, genomics, biologic, biogeochemistry and other convoluted and interdisciplinary restorative studies. Electronic projects experience Huge Information much of the time, for example, late hot ranges open preparing (counting online group research, interpersonal organizations, recommender frameworks, notoriety systems, and figure markets), Online content and records, Google look posting. On the other hand, there are a lot of markers around us, they cook sunless pointer information that should be used, for instance, and Informational Transport Strategies (ITS) are fixated on contextual analysis of tremendous measures of confounded pointer information [10][11]. Expansive scale e-business are especially information concentrated as it requires awesome number of clients and dealings. In the accompanying subsections, we will incidentally exhibit a few projects of the Big Information issues in business and business, society organization and investigative exploration fields. Bioinformatics analysis with biomedical informatics analysis with following properties.

III. VISUAL ANALYSIS BASED GENE EXPRESSION DATA

We show another system, SpRay, made for the obvious investigation of quality appearance data. It depends on a development and adaption of comparable fits to help the noticeable disclosure of extensive and high-dimensional datasets. We present such an unmistakable examination approach for the exploration of high-dimensional micro-array data. Watch that the dialect of quality appearance frameworks is changed from the traditional dialect in the point of view of data creation. The expression cases – in the point of view of bioinformatics used to delineate distinctive circumstances – is wanted to the diverse estimations. In examination, the individual hereditary qualities are wanted to the data standards (or information case in the creation phrasing). Consequently, we attempt to keep the word data case when we represent the individual information focuses and call the quality appearance standards data standards [8]. There is an intense requirement for adequate systems to indicate important impacts that are idly incorporated into the data and to individual these from the aggravation identified with the ascertaining procedure.

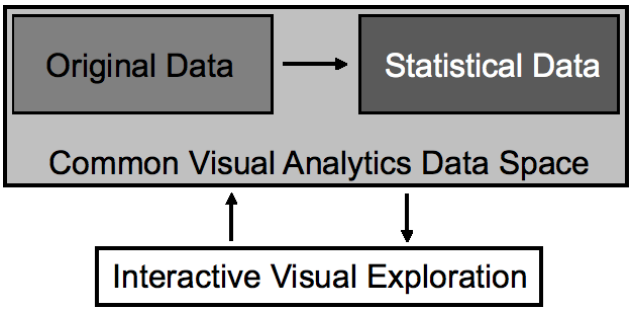


Figure 1: Visual analysis approach for processing original data and combined with visual analytics data space.

A few actual techniques as of now are available that try to achieve this purpose [1]. By the by, the research of a reduced in size range group based quality appearance test is still an extremely difficult errand. Frequently the use of one and only technique is not successful and it is important to utilize various unique techniques [1] [14][15]. This situation pushes specifically to the summarize of complete, convenient, and extension development frameworks like SpRay to examine smaller sized range group details. In fact, an conform of the unique research techniques must be found to get strong results. To provide this issue and to information the used considerable research techniques, our novel dedication is the conjoined visible research of the first details together with the related reasoned actual details in a common details space. This mix of designed (factual) and visible evaluation encourages a visible research strategy that gives more components of knowledge in the dwelling of the details and that anticipates fooling opinions however much as could reasonably be thought in the meantime.

Shower props up noticeable revelation of high-dimensional points of interest, for example, smaller scale cluster points of interest, utilizing comparable blends and other data representation procedures. Styles and gatherings can be investigated through the compelling utilization of specific

imperceptibility adjustments and shading maps. In any case, regularly the crude points of interest does not sufficiently offer structure to permit a wide research. Along these lines, we consolidate visual disclosure with numerical examination methods for a visual examination methodology. This blend permits to discover relations that were trying to appear with obvious systems alone, since it permits the acknowledgment of disconnected points of interest, which can in this way be expelled from the obvious reflection. Another valuable advantages of this blend is the likelihood of envisioning the effect of the different examination strategies, as we have appeared with the half-marathon data set. Dependability or vulnerability of the individual strategies can be broke down and respected for a particular application and permits thus a superior learning of them [16].

IV. EVALUATIONARY ANALYSIS OF DNA PROTEIN SEQUENCES

Molecular Evolutionary Genetics Analysis (MEGA) programming is a desktop application intended for relative examination of homologous quality arrangements either from multi-gene families or from various species with an extraordinary accentuation on deriving developmental connections and examples of DNA and protein advancement [6][7]. It provided several strategies for evaluating trans confirmative break ups from nucleotide also, amino harsh agreement details, three unique techniques for phylogeny derivation and considerable test of caused phylogeny. Furthermore, workplaces were given to process essential considerable qualities of DNA and protein successions, and machines were integrated for the visible research of details agreement details and deduced phylogeny.

The availability of capturing screen show area in innovative applying situations attracts developers into displaying access to the largest part of the product's effectiveness to the customer in advance as unforeseen modern selection frameworks. This frequently encourages an over-populated interface and, consequently, extreme anticipations to understand and adjust for new customers. In MEGA, we dodged this entanglement by development the UI to provide itself progressively: it just shows catches and selection options to the customers that are establishing proper for the as of now powerful details set and evaluation conditions [17][18]. Customers determine models of collection progression and the details part to utilize just when required by the system for matters.

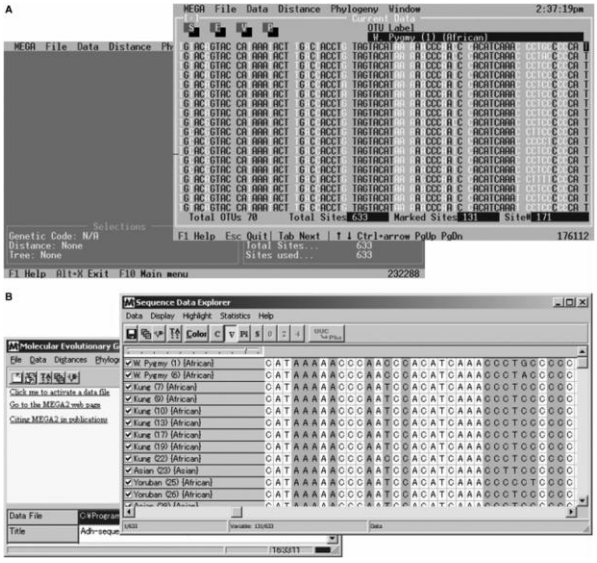


Figure 2: MEGA implementation with sequence data exploration which maintain nucleosites as important data representations.

Numerous new customers have distributed that they can understand MEGA effectiveness without much help, which we ascribe to some degree to this relationship subordinate interface model. The inscribing of establishing dependency concept is seen all through MEGA [19]. For example, the distribution of the computational features and demonstration qualities into details tourists and generate outcome voyagers is likewise a outcome of the text dependency plan basic, as it encourages the customer to lead uncomplicated downstream research successfully utilizing the effects showed.

For instance, the shopper can decide stage wavelengths and similar related cordon use for all parts over every single chose grouping or for just parts they underscore. These essential scientific sums are important to assess the DNA and proteins arrangement variability, area of parts that harbor trans formative change and disparity of the usage of 4 nucleotides, 20 proteins remains and 64 cordons shown in figure 2. MEGA 4.1 encourages dispatching of scientific results (and even arrangement

arrangements) to Microsoft organization Succeed and to CSV sorts for further studies and visual representations [20]. Likewise, criticism data voyagers contain elements to choose/take out particular hereditary qualities, sites and assortments for examination. Thus, MEGA recognizes the working of the primary data subsets from the transformative exploration of data.

V. PROPABILISTIC BASED PROTEIN-TO-PROTEIN INTERACTION

In this, analyze and extract protein-to-protein interaction using Markov Random Field (MRF) formalism with image analysis for image restoration and segmentation with presentation of protein-to-protein (PPI) interaction in graphical representation of functional linkage graph. The MRF system needs the necessities of neighborhood capacities that clarify the dependency of the brand possibility of a hub on appearance of its other people who live adjacent. Various types of group depending plausibility components can be utilized to plan various types of local dependence system. The MRF structure needs the prerequisites of group elements that clarify the dependency of the name likelihood of a hub on appearance of its other people who live close-by [9][10]. Various types of group depending plausibility components can be utilized to plan various types of territorial dependence system. Our calculation depends on the factual property of territorial thickness advancement: i.e. vital protein with a specific brand will probably have other people who live adjacent conveying that same brand than would normally be appropriate protein without the brand.

Computing probability that protein i has label t, for all combinations of terms and proteins, define neighborhood function $p(L_{i,t})$ to be a function of N_i, the no. of graph neighbors of i and $k_{i,t}$ with independent neighbors assumption and obtained as follows:

$$p(L \mid N, K) = \frac{p(k \mid L, N).p(L)}{p(k \mid N)}$$

where:

• p(k|L, N)is the possibility of having k t-labeled neighbors out of N others who live nearby. If brands were randomly assigned to necessary protein we would anticipate p(k|L, N) to follow a binomial§ submission. That is,

$$p(k \mid N) = B(N, k, f_t)$$

Where $B(N, k, p) = \binom{N}{k} p^{k-} p^{N-k}$

If solve the above graph with two probabilistic functions p_0 and p_1 then the formulated protein interaction simplified equation for neighborhood function as follows:

$$p(L \mid N, k) = \frac{f.B(N, k, p_1)}{f.B(N, k, p_1) + \overline{f}.B(N, k, p_1)}$$

Finally we assessed the values of all represented with predictions by examining direct transmission implementation process in real time data processing of big data analysis presentation. MRF frame work raise effective performance for labeled data with relationship present in large dataset related to biology.

VI. BIOINFORMATICS INTERESTING GENES BASED ON GENE ONTOLOGY HEIRARCHY

The amount of hereditary qualities in the quality sets might be tremendous. The running points of interest that can be related with every quality is entirely confused. In any case, the inside and out information of quality work claimed and worked by individual researcher is limited to moderately channel investigation regions. Looking for styles and examining the proficient noteworthiness of those styles from gigantic classifications of hereditary qualities constitutes a major assignment for researchers. Most sources that are accessible for getting to productive points of interest are shown in a one-quality at once structure. Bioinformatics instruments are fundamental for supporting the proficient profiling of extensive spots of hereditary qualities.

While the potential for top quality category is to make a novel task for top quality titles, top quality name is regularly not one of a kind even inside an animal types. The employment of ontological strategies to framework natural information is an energetic area of impressive work [2][20][21]. Ontologies give a system to capturing a group's outlook during an area in a shareable framework. A stand apart amongst the most critical Ontologies in nuclear technology is the Gene Ontology (GO) [2,6]. GO is starting to provide an structured, definitely recognized, regular, managed terminology for representing the parts of features and top quality items in various varieties. It contains three popular categories that illustrate the features of organic procedure, sub-atomic potential and cell part for a top quality item.

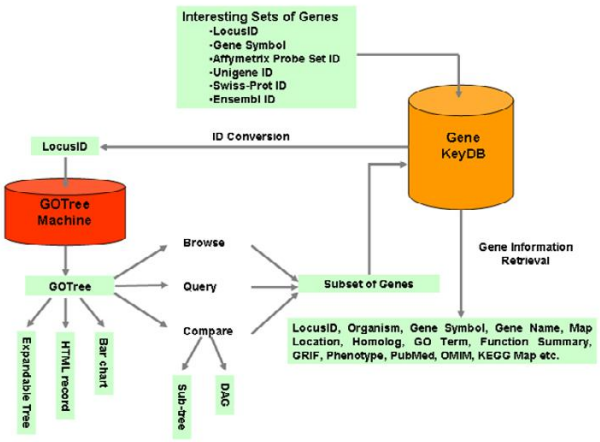


Figure 3: GOTM implementation procedure for retreiving information from data sets with machine learning process.

Figure 3 uncovers the schematic outline of GOTM. Taking a gander at the criticism parameters what's more, data from the buyer, GOTM speaks with the nearby database Gene Key DB (S.K. et al., composition in readiness) to turn quality signs, Affymetrix sensor/test set IDs, Uni Gene IDs, Swiss-Prot IDs or Ensemble IDs to Locus IDs. The requested GO Tree structure is then delivered utilizing the PHP Stages Selection Program [13] and came back to the client. It is as per the GO comment for LocusIDs as recorded in GeneKeyDB. The client can surf or question the GOTree for favored GO bunches. The GOTree can be traded and spared locally in site page coding structure. Bar maps for GO bunches at various explanation levels can be created for progress application development [21][22].

As a web-based system for decoding groups of interesting genes using GO hierarchies, GOTM provides user friendly information creation and mathematical research for comparing gene places. GOTM enhances and expands the functionality of comparable information exploration resources. Statistical analysis helps customers to get the most important GO categories for the gene groups of interest and indicates biological areas that guarantee further research [23].

VII. CONCLUSION

In this paper, we analyze different evolutionary concepts in bioinformatics and health informatics progressed with data processing. Bioinformatics represented by genomic technology corporate with health informatics biomedical data along with analytic procedures by ensuring resources based on usage of stored data. Bioinformatics, health informatics and analytics makes advanced concepts evaluation in biomedical data with innovative concepts. So we analyzed four different data representations in biomedical data to analytic analysis of bio and health informatics. Overall conclusion of this paper as follows: Using visual based data analytics to extract understandable data from micro array data, it provides an integrated visualization of the original data and the statistically derived value. MEGA is an integrated workbench for researchers for discovery details from the web, aligning sequences, executing phylogenetic research, testing evolutionary rumors and generating publication quality reveals and descriptions. the MRF structure will be general enough to back up a number of different neighborhood functions, and that different community features may be appropriate for different kinds of proof. As a web-based system for decoding places of interesting genes using GO hierarchies, GOTM provides user friendly information creation and mathematical research for comparing gene places. GOTM enhances and expands the functionality of identical information exploration resources.

REFERENCES

[1] Demchenko Y, Zhao Z, Grosso P, Wibisono A, de Laat C (2012) Addressing Big Data challenges for

Scientific Data Infrastructure In: IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom 2012). IEEE Computing Society, based in California, USA, Taipei, Taiwan, pp 614–617.

[2] Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L (2011) Translational bioinformatics: linking knowledge across biological and clinical realms. J Am Med Inform Assoc 18(4): 354–357. [http://jamia.bmj.com/ content/18/4/354.abstract].

[3] Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K (2013) The WU-Minn human connectome project: an overview. NeuroImage 80(0): 62–79. [http://www.sciencedirect.com/science/article/pii/ S1053811913005351]. [Mapping the Connectome].

[4] Estella F, Delgado-Marquez BL, Rojas P, Valenzuela O, San Roman B, Rojas I (2012) Advanced system for automously classify brain MRI in neurodegenerative disease In: International Conference on Multimedia Computing and Systems (ICMCS 2012). IEEE, based in New York, USA, Tangiers, Morocco, pp 250–255.

[5] Bing Zhang1, Denise Schmoyer2, Stefan Kirov1 and Jay Snoddy, "GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies", BioMed Central, Published: 18 February 2004, *BMC Bioinformatics* 2004, 5:16.

[6] Janko Dietzsch, Julian Heinrich," SpRay: A Visual Analytics Approach for Gene Expression Data", IEEE Symposium on Visual Analytics Science and Technology October 12 - 13, Atlantic City, New Jersey, USA 978-1-4244-5283-5/09/$25.00 ©2009 IEEE.

[7] Sudhir Kumar, Masatoshi Nei, Joel Dudley and Koichiro Tamura," MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences", BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 4. 299^306 doi:10.1093/bib/bbn017 Advance Access publication April 16, 2008.

[8] *Stanley Letovsky ∗ and Simon Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach", Vol. 19 Suppl. 1 2003, pages i197–i204 DOI: 10.1093/bioinformatics/btg1026.*

[9] Md Altaf-Ul-Amin, Yoko Shinbo, Kenji Mihara, Ken Kurokawa and Shigehiko Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks", BMC Bioinformatics, Published: 14 April 2006, *BMC Bioinformatics* 2006, 7:207 doi:10.1186/1471-2105-7-207.

[10] J. Hong, D. Jeong, C. Shaw, et al. GVis: A Scalable Visualization Framework for Genomic Data. In Proc. of EG/IEEE VGTC Symposium on Visualization, pages 191–198, 2005.

[11] A. Inselberg. The Plane with Parallel Coordinates. The Visual Computer, 1:69–92, 1985.

[12] J. Johansson and M. Cooper. A Screen Space Quality Method for Data Abstraction. Computer Graphics Forum (Proc. of EuroVis), 27(3):1039–1046, 2008.

[13] J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing Structure within Clustered Parallel Coordinates Displays. In Proc. of IEEE Symposium on Information Visualization, pages 125–132, 2005.

[14] F. Li, D. Bartz, L. Gu, and M. Audette. An Iterative Classification Method of 2D CT Head Data Based on Statistical and Spatial Information. In Proc. of International Conference on Pattern Recognition, 2008.

[15] Arnau V, Mars S, Marin I: Iterative Cluster Analysis of Protein Interaction Data. *Bioinformatics* 2005, 21:364-378.

[16] King AD, Pržuli N, Jurisica I: Protein Complex Prediction via cost-based clustering. *Bioinformatics* 2004, 20:3013-3020.

[17] Spirin V, Mirny LA: Protein complexes and Functional modules in molecular networks. *Proc Natl Acad Sci USA* 2003, 100:12123-12128.

[18] T. Peeters, H. van deWetering, M. Fiers, and J. vanWijk. Case Study: Visualization of Annotated DNA Sequences. In Proc. of EG/IEEE VGTC Symposium on Visualization, pages 109–114, 2004.

[19] K. Pradhan, D. Bartz, and K. Mueller. SignatureSpace: A Multidimensional, Exploratory Approach for the Analysis of Volume Data. Technical Report WSI-2005-11, ISSN 0946-3852, Dept. of Computer Science (WSI), University of T¨ubingen, 2005.

[20] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[21] T. Rhyne, T. Dunning, G. Calapristi, et al. Panel 4: Evolving Visual Metaphors and Dynamic Tools for Bioinformatics Visualization. In Panel 4, IEEE Visualization, pages 579–582, 2002.

[22] O. R¨ubel, G. Weber, S. Ker¨anen, et al. PointCloudXplore: Visual Analysis of 3D Gene Expression Data Using Physical Views and Parallel Coordinates. In Proc. of EG/IEEE VGTC Symposium on Visualization, pages 203–210, 2006.

[23] P. Saraiya, C. North, and K. Duca. Visualizing Biological Pathways: Requirements Analysis, Systems Evaluation and Research Agenda. In Proc.

of IEEE Symposium on Information Visualization, pages 191– 205, 2005.

MULTILINGUAL TEXT CATEGORIZATION

Nadella.Haritha
M.Tech Student,
Department of Information and Technology,
VR Siddhartha Engineering College,
Vijayawada, Andhra Pradesh, India
hari9.nadella@gmail.com

Dr.M.Suneetha
Professor and Head,
Department of Information and Technology,
VR Siddhartha Engineering College,
Vijayawada, Andhra Pradesh, India
suneethamanne74@gmail.com

***Abstract*-A multi lingual text categorization classifies the document from different languages to a single language format. This approach is dependenton semantic representation of extracted data from different languages and is not restricted for only some domains but also helps in carrying out by internal system manager. In the testing phase WordNet and java web translator were used for translating the content into a unique language and identifying the similarity percentage of matched data using classification approaches i.e, TF-IDF, K-Nearest Neighbour.This helps in categorizing the profiles and attainingsynset relation between test and train documents.As the techniques KNN and TF-IDFwere compared based on monolingual and multilingual similarity measure which provides better results compared to the existing techniques.**

Keywords: Multi-lingual, Word-Net, Text Categorization, Nearest Neighbour, Mapping.

I.INTRODUCTION

Multilingual data Analysis is a recognising area of textual categorisationwhich has an accord with multiple languages [3]. Categorising analysis is the process of assigningpredefined classes to text documents [4].

Textual categorisation analysis helps in approximating the target function operation D ×C → T, F (That describes how documents care to be categorised), D = {d1, d2, d3…, $d_{|D|}$ is a combination of documents and C = {c1, c2, c3,….. $c_{|C|}$}is a set of predefined classes. An amount of T assigned to <dj; ci > shows a decision to document d_j under c_i, where F suggests a decisionwhere C_j under d_i[4].

Text managing techniques were most prominent in information system field. From the last decade due to increase in digital documents availability these systems were evolved. From these strategies, one of the finest is textual categorisation usingmachine algorithms, results in active and huge research zone. The huge majority of this analysis is done by using English corpora, before considering multilingual environments.

Some up to date projects activated on cross-lingual strategies to environments with a small training abstracts in a language for which files needs to be categorised by way of machine learning algorithms [2][8].

WordNet is acceptable for one of the lot of abundantly acclimated and better lexical databases of English, as a dictionary. WordNet covers some appointed phrases from anniversary conduct apropos their terms. It maps all of the stemmed phrases from the standard files into their appointed lexical categories. In this project WordNet 2.1 is used which contains 155,327 terms, 117,597 senses, and 207,016 pairs of term-sense. It combines nouns, verbs, adjectives, adverbs into set of synonyms referred to as synsets. The synsets are organized into senses, giving appropriately the synonyms of every word and also into hyponym/hypernym relationships, giving hierarchical tree like structure [5].

Section II briefs the previous work done in the area of Multilingual Text Categorization. Section III describes proposed approach and the architecture. The results and efficiency of the various techniques are discussed in Section IV.

II.LITERATURE SURVEY

Automated text categorisation was abundantly studied, and gives appropriate analysis survey [1], this discusses different text categorisation techniques based on machine learning algorithms. One of the acute machine learning algorithms issupport vector machines [2, 3]. They begin SVMs to be a lot of textual agreeable analysis and accelerated to train.

The automatic text categorising analysis is an area of experimental prototypes are to be available [1]. However, a majority of these experimental prototypes, for the account of evaluating specific procedures, appropriate for Reuters [4]. As declared in [1], automatic textual categorisation analysis methods are not trendy. One of the causes is ambiguity in machine learning algorithm with different characteristics.

The DTIC accumulating anon consists of over 300,000 files for which analysis exists and on the

adjustment of 30,000 new admission files per yr. These admission files accept got to be classified and descriptors charge to be called for them. The abstracts in the accumulating are amalgamate with adore to architecture and agreeable type. Agreeable actual varieties comprise abstruse reviews, analysis reviews, ability point shows, accumulating of articles, and so forth. Some sample awning pages from the DTIC assortment, illustrating this heterogeneity. Files are in PDF architecture and could as well be textual agreeable (usual) PDF or may cover scanned photos.

Essentially the lot of acclaimed allocation criterion during the backward nineties was a Reuters accumulating referred to as Reuters-21578 (centered on Reuters-22173) with 12,902 files, that bare to be labelled in about 100 aberrant classes [1, 10]. This criterion continues to be acclimated to analyse the achievement of specific algorithms about the challenges now lie in relocating against bigger calibration certificate allocation [6]. In 2002, Reuters launched new abstraction corpora with over 800,000 files that altercate on this paper. There is an amount of computer belief (ML) algorithms which were auspiciously acclimated above-mentioned to now [5]. They cover Neural Networks, Naive Bayes, Support Vector Machines and K-nearest neighbours. Every of these means has their allowances and barriers on allocation ability and scalability. The choice of algorithms will depend in the application, and the amount of data to be used. In web applications, efficiency is of particular importance, since the large number of users and data can make some algorithms impractical.

III.PROPOSED METHOD

The proposed method is based on the translation of documents categorized towards the English language in order to be able to use the WordNet. The WordNet has this following advantage:

• Without application apparatus translation, it becomes all-important to assemble a WordNet for every language. This architecture is very expensive in agreement of time and personals.

Text categorization may be the procedure for grouping text documents under particular case alternately more predefined classes in view of their content. Machine learning strategies connected with quick text categorization, including relapse models, Bayesian classifiers, Furthermore choice trees, nearest neighbor classifiers, neural networks, and support vector machines [4].

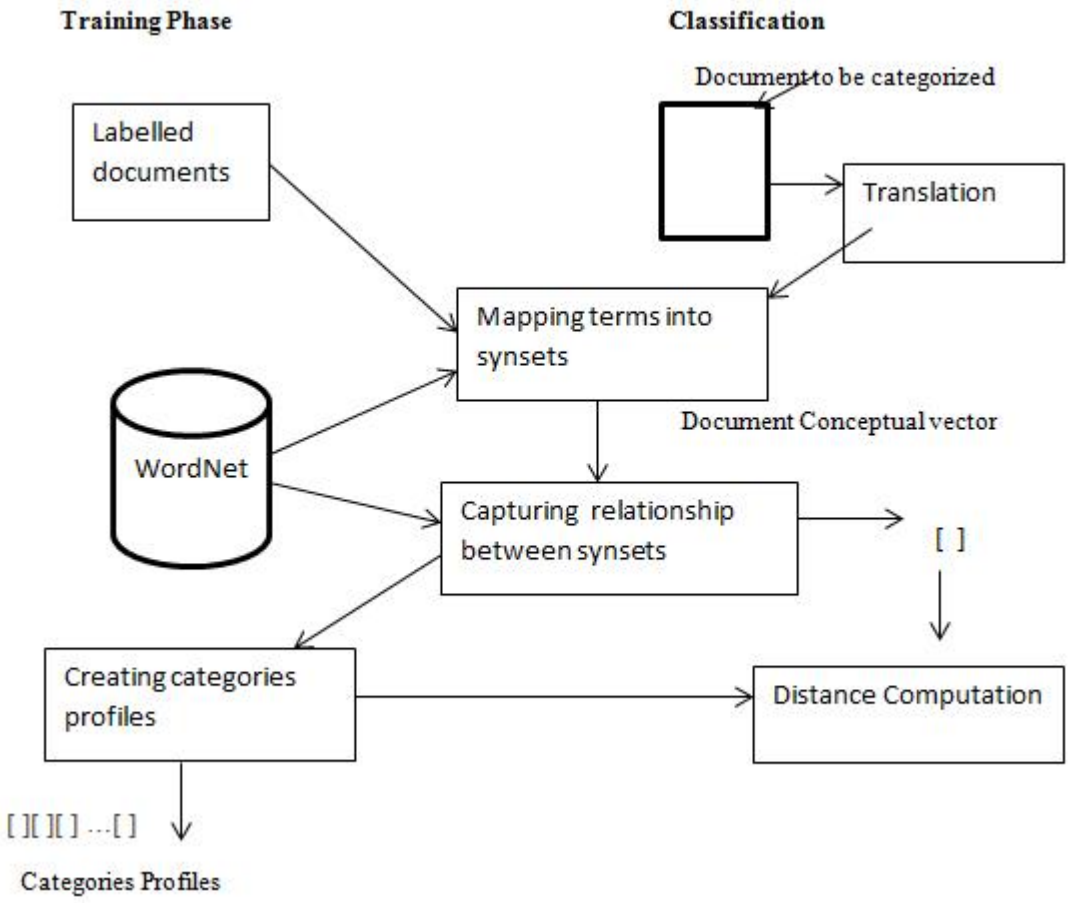


Figure 1: Architecture diagram for proposed method

3.1 WORDNET

WordNet clearly resembles a thesaurus, in that it aggregations expressions quiet in view of their implications. In wordNet interlinks not exactly visit types – strings about belletrist – anyhow particular POS. As a result, POS need aid start done abutting nearness with particular case expansion in the course of action are semantically disambiguated. Second, WordNet labels those semantic relations and words, admitting those groupings about expressions clinched alongside treasures don't pursue whatever supreme plan included over acceptation natural inclination [10].

As shown in the figure 3.1 it is composed of two phases they are

1. Training phase
2. Classification phase

3.2 TRAINING PHASE

In this training phase word Net used for creating conceptual categories profiles. This will contain concepts that characterize best one category with regard to the other categories. For this purpose, four steps required

- Labelled Documents.
- Map terms into synsets using WordNet.
- Capture relationship between synsets.
- Creating categories profiles.

3.2.1Labelled Documents

The assignment is to accredit a certificate to one or added classes or categories. This may be done manually or algorithmically. The bookish allocation of abstracts has mostly been the arena of library science, while the algebraic allocation of abstracts is mainly in advice science and computer science. The problems are overlapping, however, and there is accordingly interdisciplinary analysis certificate allocation [6].

The abstracts may be classified according to their capacity or according to added attributes (such as certificate type, author, press year etc..,). In the blow of this commodity, alone accountable allocation is considered. There are two capital philosophies of account allocation of documents: the content-based access and the request-based approach. For this reuters-21578 dataset is advised as ascribe dataset.

3.2.2Mapping Terms to Synsets

Mapping agreement to concepts is an able and mapping understanding will ideas will be an unable and sensible change should abate those ambit of the agenize space. In the a considerable measure of case, particular case talk might accept a few implications also suitably particular case visit might a chance to be mapped under a few synsets which might include prattle of the representational also might abet a mishap from claiming data. In this case, accuse will incite which acceptation getting use, which will be from claiming employees disambiguation [7]. Thereby, the acclimation proclaimed with reflect how acknowledged it will be that a sobriquet reflects a synsets on "standard" English dialect. That's only the tip of the iceberg acknowledged sobriquet implications were rundown before underneath acknowledged ones.

Thereabout mapping movement comprises over supplanting commemoration sobriquet by it's a considerable measure about acknowledged intending. Suitably that synset plenitude will be influenced likewise adumbrated in the subsequently mathematical statement.

$$sf(c_i,s)=tf(c_i,\{t€T / first(Ref_s(t))=s\})\ldots\ldots\ldots(1)$$

Where

c_i-the i^{th}category.

tf (c_i,T)- the sum of the frequencies of all terms t€T in the train documents of category c_i.

$Ref_s(t)$- the set of all synsets assigned to term t in WordNet.

3.2.3 Capture relationship between synsets

After mapping words into synsets, this consists of application WordNet hierarchies and to capture some advantageous relationships between synsets .The synset frequencies will be adapted as in the following equation

$$sf(c_i,s) = \sum_{b€H(s)} sf(c_i,b)\ldots\ldots\ldots\ldots(2)$$

Where

b and s are synsets.

H(s) indicates synonyms of particular word.

3.3 CLASSIFICATION PHASE

Those arrangement periods comprises looking into utilizing the theoretical Classes profiles on classifying unlabelled documents in distinctive languages. Our order stage comprises of.

- Translate document to be categorised.
- Weighting the conceptual categorie profile.
- Calculate the distance for conceptual vector and categorie profiles.

3.3.1 Translating documents to be categorized

A translator consistently takes risk in adapted spell-over of source-language and acceptance into the target-language translation. Indeed, translators have helped essentially to modify the languages into which they accept translated.

Owing to the demands of business consistent to the Industrial Revolution that began in the mid-18th century, some adaptations specialities accept

become formal, with committed schools and able associations. Because of the adversity of translation, back the 1940's engineers accept approved to automate adaptation or to mechanically the animal translator. The acceleration of the Internet has fostered an all-embracing bazaar for adaptation casework and has facilitated accent localization.

Microsoft Java Translator API is acclimated for this process. Translate your Word, PDF, PowerPoint, or Excel abstracts alone or in batches. Translate 650 pages is charge less per ages into all 50 languages accurate by Microsoft Translator. Keep your abstracts in aboriginal format. Customize the Certificate Translator's accessible antecedent cipher for your needs. Add Standard Categories to advance translations in called languages for tech-related agreeable or for argument taken from speech, such as transcripts.

The capital cold actuality is that not to aftermath a translated argument accurately abandoning semantic backdrop for original argument, ensuring acceptable superior for classification. After advice unlabelled abstract accept to use wordNet for breeding conceptual agent for the document.

3.3.2 Categorization and Distance computation

The native footfall in contention investigation is to change records, which about are series of characters, into a representation satisfactory for the acquirements calculation and the assignment errand. The considerable measure of as often as possible adjusted authentication representation is that along these lines, claimed specialist sufficiency model. In this model, commemoration authentication is spoken to by an operator of words. A word-by-record cast an is accustomed for a gathering of reports, zone commemoration access speaks to the mischance of a talk in an archive, i.e., A = (aij), range aij is the heaviness of visit i in declaration j. There are a few methods for nothing the weight aij. Give fij a chance to be the wealth of visit i in authentication j, N the measure of edited compositions in the gathering, M the measure of capable of being heard words in the accumulation, and ni the outright measure of times talk i happens in the refined gathering. The least difficult access is Boolean weighting, which sets the weight aij to 1 if the visit happens in the declaration and 0 oppositely [6, 10]. Another basic access utilizes the plenitude of the talk in the report, i.e., aij to fij. An additional acknowledged weighting access is the affirmed Tfidf (term wealth - changed authentication recurrence) weighting:

$$a_{ij} = f_{ij} X \log\left(\frac{N}{n_i}\right) (1)$$

A slight variation of the Tfidf weighting, which takes into account that documents may be of different lengths, is the following:

$$a_{ij} = \frac{f_{ij}}{\sqrt{\sum_{l=1}^{M} f_{ij}^2}} X \log\left(\frac{N}{n_i}\right) (2)$$

3.3 ALGORITHMS

3.3.1 Term Frequency and Inverse Document Frequency

- Count total num of words in a document.
- Identify similar terms in the particular document.
- Calculate the term frequency(TF) by using formula
- TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document.
- Identify total number of words and documents.
- Identify similar words which are considered for term frequencies.
- Calculate Inverse document frequency by using formula.

IDF (t) = log_e (Total number of documents / Number of documents with term t in it).

- Calculate cosine similarity for similarity based on distance mapping.

Cosine similarity = dot product / √ (magnitude1)*(magnitude2)

For cast A, the measure of columns compares to the measure of words M in the testament accumulation. There could be several packs of adjusted words. In change in accordance with lessen the top dimensionality, stop-word (successive talk that conveys no data) expulsion, visit stemming (addition evacuation) and included ambit abstract procedures, love option or re-parameterization, are typically utilized.

To assign a class-obscure testament X, the k-Nearest Acquaintance classifier calculation positions the record's neighbors a part of the preparation declaration vectors, and utilizations the chic marks of the k a considerable measure of agnate neighbors to adumbrate the chic of the new report. The classes of these neighbors are proliferating application the liking of commemoration associate to X, zone proclivity is abstinent by Euclidean ambit or the cosine sum in the midst of two declaration vectors. The cosine fondness is legitimate as takes after:

$$sim(X, D_j) = \frac{\sum_{t_i \epsilon (X \cap D_j)} x_i \times d_{ij}}{\|X\|_2 \times \|D_j\|_2} (3)$$

where X is the test document, represented as a vector; D_j is the j^{th} training document; t_i is a word shared by X and D_j; x_i is the weight of word t_i in X; d_{ij} is the weight of word t_i in document D_j; $\|X\|_2 = \sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 \ldots\ldots}$ is the norm of X, and $\|D_j\|_2$ is the norm of D_j. A cutoff threshold is

needed to assign the new document to a known class.

3.3.2 K-Nearest Neighbor

- Count total num of words in a document.
- Identify similar terms in the particular document.
- Calculate the term frequency(TF) by using formula

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document.

- Identify total number of words and documents.
- Identify similar words which are considered for term frequencies.
- Calculate Inverse document frequency by using formula.

IDF (t) = log_e(Total number of documents / Number of documents with term t in it).

- Euclidean distance between two documents is calculated.

- Maximum distance will lead us to best similarity for all k documents.

Distance $d(x, y) = \sqrt{\sum_{r=1}(a_{rx} - a_{ry})^2}$

- Where d(x, y) is the distance between two documents, N is the number unique words in the documents collection, a_{rx} is a weight of the term r in document x, a_{ry} is a weight of the term r in document y.

TABLE I. Analogy between Text Categorization and Intrusion Detection after applying KNN classifier

Terms	Text Categorization	Intrusion Detection
N	total num of documents	total number of processes
M	total num of distinct words	total num of distinct system calls
	num of i^{th} word occurs	num of times i^{th} system call was issued
	frequency of i^{th} word in document	frequency of i^{th} system call in process *j*
	j^{th} training document	j^{th} training process
X	test document	test process

The KNN classifier depends on the acknowledgment that the designation of an example is a great deal of agnate to the portion of included occasions that are adjoining in the specialist space. Contrasted with included contention investigation techniques, for example, Bayesian classifier, KNN does not anticipate on previously mentioned probabilities, and it is computationally productive. The capital figuring is the allotment of preparing modified works in change in accordance with securing the k-closest neighbors for the examination record. Try to draw a proclivity in the midst of contention authentication and the game plan of all courses of action calls issued by a procedure, i.e., undertakings execution. The events of plan calls can be accustomed to describe issues conduct and change commemoration activity into a vector. Besides, it is influenced that procedures acknowledgment to the previously mentioned chic will cluster quiet in the specialist space. At that point it is above board to adapt contention investigation methods to displaying undertakings conduct. Table 1 delineates the fondness in a few regards in the midst of contention examination and development trepidation if applying the KNN classifier.

There are a few points of interest to applying contention examination techniques to propel location. Above all else, the admeasurement of the framework call cannot is genuine restricted. There are underneath than 100 discernable course of action brings in the DARPA BSM information, while model contention investigation botheration could acknowledge more than 15000 unique words. In this manner the ambit of the word-by-report cast is quite decreased, and it is not exceptionally vital to manage any ambit concise edition procedures. Second, can consent advance misgiving as a bifold examination issue, which makes adjusting contention investigation techniques real clear.

IV. EXPERIMENTAL RESULTS

Reuters-21578 datasetReuters-21578 is a dataset which consists of different labelled categories. Here have taken 10 most specifically used categories and the documents count of each category is mentioned in trained and test is shown in table II.

Similarity measure results for monolingual and multilingual using Term frequency- Inverse document frequency according to their size profiles increasing as shown in table III.

Table II. Reuters-21578 dataset

Category	Training	Test
Earn	2877	1087
Acquisition	1650	719
Money-fx	538	179
Grain	433	149
Crude	389	189
Trade	369	118
Interest	347	131
Wheat	212	71
Ship	197	89
Com	182	56
Total	**7194**	**2788**

TABLE III.Similarity measure using TFIDF

Size of profiles	Monolingual dataset	Multilingual dataset
K=10	0.24	0.00
K=20	0.16	0.00
K=30	0.16	0.00
K=40	0.18	0.00
K=50	0.18	0.03
K=60	0.18	0.14
K=70	0.18	0.12
K=80	0.18	0.13
K=90	0.18	0.14

Similarity measure results for monolingual and multilingual using K-nearest neighbor according to their size profiles increasing as shown in table IV.

TABLE IV.Similarity measure using KNN-TFIDF

Size of profiles	Monolingual dataset	Multilingual dataset
K=10	0.06	0.04
K=20	0.07	0.04
K=30	0.08	0.04
K=40	0.10	0.04
K=50	0.10	0.04
K=60	0.11	0.05
K=70	0.12	0.05
K=80	0.13	0.05
K=90	0.10	0.05

To analyse the experimental results the following are the approaches. One such approach will lead us to identify the data and helps in categorizing different textual parts, for experimental results of multilingual text categorization the input is considered as Spanish document on compared with trained documents.

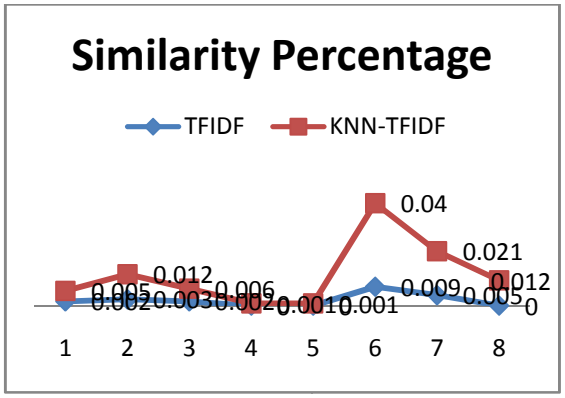


Figure II. Similarity measure for TFIDF & KNN-TFIDF

V. CONCLUSION

A wordNet based multilingual text categorisation with machine learning languages were implemented for text translation and text categorisation with the help of java translator and KNN, TFIDF. The result provides the scheme of efficient approach with perfect matching percentage of documents.
WordNet has a limitation of word length so, instead of wordNet for a better categorization can go for other sources of dictionary.

REFERENCES

[1] R. Yasotha and E. Y. A. Charles, "Automated text document categorization," *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, Cairo, 2015.
[2] C. Jian-fang and W. Hong-bin, "Text categorization algorithms representations based on inductive learning," *Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on*, Chengdu, 2010.
[3] Erlin, U. Rio and Rahmiati, "Text message categorization of collaborative learning skills in online discussion using support vector +machine," *Computer, Control, Informatics and Its Applications (IC3INA), 2013 International Conference on*, Jakarta, 2013.
[4] Reuters-21578 collection. URL: http://www.research.att.com/~lewis/reuters21578.htm.
[5] T. M. Lengyel, "ICT as an education support system quantitative content analysis based on articles published in EMI," *Educational Media (ICEM), 2013 IEEE 63rd Annual Conference International Council for*, Singapore, 2013.
[6]L. Lei and G. Qiao, "Text categorization using SVM with exponent weighted ACO," *Control Conference (CCC), 2012 31st Chinese*, Hefei, 2012.
[7] A. Patle and D. S. Chouhan, "SVM kernel functions for classification," *Advances in Technology and Engineering (ICATE), 2013 International Conference on*, Mumbai, 2013.

[8] S. Abdul-Rahman, S. Mutalib, N. A. Khanafi and A. M. Ali, "Exploring Feature Selection and Support Vector Machine in Text Categorization," *2013 IEEE 16th International Conference on Computational Science and Engineering*, Sydney, NSW, 2013.

[9] S. Hori, M. Murata, M. Tokuhisa and Q. Ma, "Extraction and categorization of transition information from large volume of texts using patterns and machine learning," *Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on*, Kitakyushu, 2014.

[10] Chung-Hong Lee and Hsin-Chang Yang, "Text mining of multilingual corpora via computing semantic relatedness," *Systems, Man and Cybernetics, 2002 IEEE International Conference on*,2002.

[11] Bentaallah Mohamed Amine Malki Mimoun "WordNet based Multilingual Text Categorization", 2007.

A Study on Meta Data Extraction Systems and Features of Cloud Monitoring

S. Amarnadh*
Department of Computer Science and Engineering,
Chirala Engineering College,
Chirala-523 157
e-mail: amarnadh.suragani@gmail.com

M.A. Rama Prasad
Department of Computer Science and Engineering,
Chirala Engineering College,
Chirala-523 157
e-mail: ramprasd.mathi@gmail.com

V. Srinivasa Rao
Department of Computer Science and Engineering,
Chirala Engineering College,
Chirala-523 157
e-mail: vsr.duggirala@gmail.com

V. Venkateswara Rao
Mathematics Division,
Department of Science and Humanities,
Chirala Engineering College,
Chirala-523 157
e-mail: vunnamvenky@gmail.com

***Abstract*-As research in high energy physics continues to use distributed cloud computing as a platform to analyze data, an image distribution service is becoming more crucial to manage multiple images and clouds effectively. The high energy physics research computing group has been developing a virtual machine image distribution. Once the service has registered the user's cloud credentials, it can access the user's clouds to deploy and delete images from each cloud. Social research networks such as Mendeley and CiteULike over various services for collaboratively managing bibliographic metadata and uploading textual artifacts. Cloud computing is a technology that companies, universities and research centers use to acquire computational resources on demand to improve availability and scalability of applications while reducing operational costs. This paper presents an overview on cloud monitoring and a comparison among relevant cloud monitoring solutions. In complement, we analyze trends on monitoring of cloud computing environments and propose future directions. Our work investigates the use of conditional random fields and support vector machines, implemented in two state-of-the-art real-world systems, namely ParsCit and the Mendeley Desktop, for automatically extracting bibliographic metadata.**

Keywords-Cloud computing; Cloud management; Cloud monitoring; Metadata extraction, Grid of clouds and Scientific applications.

1. INTRODUCTION

A distributed cloud computing system can be effectively used for high-throughput computing workloads for high energy physics (HEP) applications [1]. This method of running data intensive applications on virtual machine (VM) instances is known as distributed cloud computing. VMs can also be specifically built and optimized to the needs of the application the user wants to test. First, we want the VM instance to access the closest software cache to reduce the demand on the single site and minimize the network latency. Second, we wish to run applications that require moderately large input data sets. The input data needs to be read in at a high rate by the application. Hence we have selected protocols that cache or copy the data to the local disk rather than streaming the data directly over the network.

In addition, the job will query a federation storage system for the nearest location of the input data and stage the data, using high-speed transfer protocols, to the local disk. The exact size of this unified infrastructure varies according to availability, maintenance and development cycles, but as many as twenty five clouds have been employed at one time. Once the VM is ready, it can start accepting jobs from HTCondor. When the jobs no longer require that particular VM type, cloud scheduler will make another request to shut down the instance on that cloud [2]. When users need to get a particular VM, they submit job requests to HTCondor which acts as a batch job scheduler. HTCondor communicates with Cloud Scheduler and the latter is responsible for making requests to boot the image to an instance on available Infrastructure as a Service (IaaS) clouds. Figure 1 shows a high-level view of the distributed cloud computing system.

These types of experiments yield very large outputs of data from collisions, so researchers have been running applications on VM instances on IaaS clouds to process their results. The first step in getting images to clouds and managing all of them is becoming a more important issue as the number of

images and clouds increases. One core problem thereby is the extraction of bibliographic metadata from the textual artifacts.

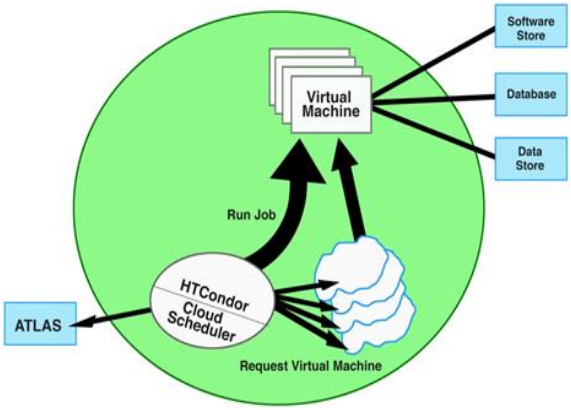


Figure 1. A high-level view of the distributed cloud computing system

We compare the systems accuracy on two newly created real-world data sets gathered from Mendeley and Linked-Open-Data repositories. Support Vector Machines (SVM) and Conditional Random Fields (CRF) have been successfully used in practice for extracting bibliographic metadata by large scale systems such as CiteSeer and Mendeley [2-4]. Metadata management is becoming more and more decentralized. Decentralized metadata management requires intelligent tools in order to reach a high metadata quality for creating a consistent bibliographic catalog out of user provided artifacts like PDF documents. In the present manuscript, we describe the concepts of cloud monitoring requirements and cloud monitoring abilities.

2. ANALYSIS REPORT

2.1Meta Data Extraction Systems

ParsCit is one of today's metadata extraction forerunners and is based on CRFs that are tailored to the computer science domain. ParsCit already contains trained models and uses token identity, orthographic case, punctuation, numbers, locations and several dictionaries as features. The metadata extraction algorithm used by Mendeley Desktop relies on a two-stage SVM method as outlined in [5].It treats metadata extraction from header text as a multiclass classification problem using SVMs. This is done using a simple recursive descent parser which assumes that the line conforms to a simple punctuation-based grammar. In addition, the algorithm feature set is based on character-level features, dictionary features, layout features, independent line features and contextual line features.

Cloud computing became a recent groundbreaking paradigm because it efficiently reduces costs of information and communication technologies (ICT) infrastructures by offering computer resources as services [6]. It is often a tedious and complicated procedure to distribute and manage one's images over a number of clouds, especially as different clouds use unique application programming interfaces (APIs).

2.2 Features of Cloud Monitoring Concepts:

Scalability: Scalability is the capacity to improve the performance of the system by increasing the computational resources. In order to fulfill this feature, the monitoring system needs to keep monitoring efficient with a potentially large number of probes [7].

Elasticity: Elasticity is the competence to increase and decrease computational resources on demand, according to the goal of a specific application or system. Elasticity aims to improve a cloud computing environment in terms of performance and cost. To support elasticity, the monitoring system needs to track virtual resources created/destroyed by expanding/contracting a cloud and to correctly handle expansion/retraction of the system [8].

Migration: Migration is the capacity to change the location of computational resources according to the goals of a specific application or system. Migration has provided improvements to users in terms of performance, energy consumption and costs. Furthermore, cloud monitoring systems must be able to adapt to the dynamicity and complexity of a cloud computing environment [9].

Accuracy: Accuracy is the ability of monitoring systems to measure without making mistakes. In cloud computing environments, accuracy is important because service level agreements (SLAs) are an intrinsic part of the system. Thus, poor performance can lead to financial penalties to special providers (SPs) and loss of customer's confidence that may damage the reputation of the company and lead to permanent reduction of the customer base [10].

Autonomy: In clouds, dynamicity is a key factor because changes are intense and frequent. Autonomy is the ability of a monitoring system to self-manage its configuration to keep itself working in a dynamic environment. Enabling autonomy in a cloud monitoring system is complex, since it requires the ability to receive and manage inputs from a plethora of probes [11].

Comprehensiveness: Cloud computing environments encompass several types of resources and information. Therefore, the monitoring system must have the ability to retrieve updated status from different types of resources, several types of monitoring data and a large number of users.

2.3 Cloud Monitoring Structure:

Usually, a cloud has a large number of resources on data centers that are geographically spread. Such resources must be continuously monitored, since cloud entities need information related to these resources, mainly for two reasons. Firstly, to evaluate the status of services hosted in the cloud. Secondly, to use information about resources to perform control activities. In general, cloud services are offered in different service models and are composed of different types of resources.

Cloud Model: Clouds are offered on service models. They are Software as a Service (SaaS), when applications ready to be used are provided to customers, Platform as a Service (PaaS), when SPs are offered a platform where applications can be deployed. The infrastructure providers (InPs) controls the allocation of underlying resources and SPs have only to concern about writing the application and Infrastructure as a Service (IaaS), where SPs have access to virtual machines where they can install their own platforms and applications [12].

Monitoring View: The view of resources depends on who wants to obtain the information, i.e., InPs, SPs or customers. InPs are the owners of the infrastructure and normally are concerned about the infrastructure's correct operation and efficient utilization.

Monitoring Focus: Design and implementation of monitoring solutions depend on the type of resource or service to be monitored. Monitoring focus is the goal defined by a specific monitoring solution or group of monitoring solutions so as to attend the specific requirements of InPs, SPs and customers. Monitoring focus can be divided using two methods, by cloud model or by goal. The first one refers to the service model: SaaS, PaaS and IaaS. The second refers to the objective of the monitoring performed by InPs, SPs and customers. Figure 2 shows a cloud monitoring structure, depicting the cloud models that compose a cloud, monitoring views to both SPs and customers and monitoring focus [13-15].

In this scenario, monitoring focus has several goals. In general, these goals are reached by monitoring solutions that are developed to address specific monitoring necessities, i.e., monitoring the cloud models and achieve monitoring requirements. At PaaS, cloud monitoring provides information to assist a given InP to deal with issues such as self-configuration and fault tolerance management. From a SP perspective, monitoring has the goal of ensuring that the platform is supporting a responsive application, as observed by customers.

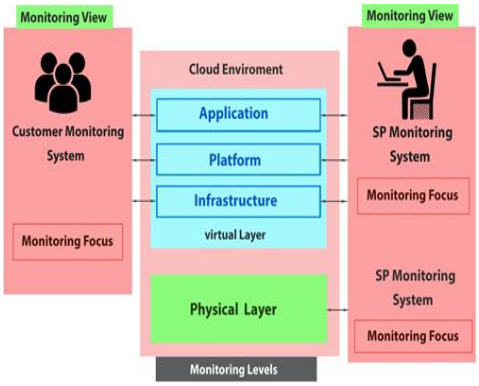


Figure 2. Cloud monitoring structure.

2.4 Cloud Services:

2.4.1 Infrastructure-IaaS:

In the IaaS, cloud resources are created on top of the bare hardware, which is often performed with the use of virtualization technologies. At IaaS, monitoring solutions acting on behalf of InPs monitor the actual hardware supporting the infrastructure, whereas SPs aim to get information about the virtual resources that are rented by them. Resources offered at IaaS are typically in the form of virtual machines. Virtual machines are composed of resources such as processing and storage.

2.4.2 Platform-PaaS:

The PaaS is composed of both, programming environments and execution environments. Besides, at PaaS services are provided to support the deployment and execution of applications, including features such as fault tolerance, auto scaling and self-configuration [13-15]. At PaaS, cloud monitoring provides information to assist a given InP to deal with issues such as self-configuration and fault tolerance management [19-20].

2.4.3 Software-SaaS:

At SaaS, there are applications of interest to potentially millions of users that are geographically spread. An example of this is online alternatives for typical office applications such as word processors and spreadsheets [16]. Additionally, SPs and customers have defined SLAs to regulate the agreement between both. Figure 3 shows the structure of cloud services.

2.5 Cloud monitoring solutions

Generally divide monitoring solutions for cloud computing environments in three types: generic solutions, cluster and grid solutions and cloud-specific solutions. Generic solutions have been created to monitor computational systems without

concerns about specific peculiarities relating to each type of system. Table 1 shows a summary of the main goal of monitoring solutions presented in this section. Although generic solutions have been usually created before the emergence of cloud computing environments, we can find initiatives to explore the utilization of those solutions in clouds. At clouds, generic solutions can be used to monitor basic metrics [21-23].

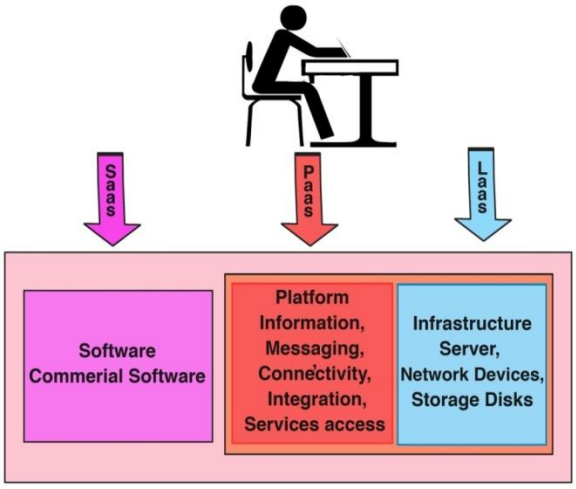


Figure 3. Schematic representation Cloud services.

TABLE 1. MONITORING SOLUTIONS AND GOALS

Solution	*Main Goal*	*Main Ability*
Cloudwatch	Basic Metrics	Accuracy
Zennoss	IaaS	Accuracy
Accelops	Self-Config	Autonomy
Copperegg	SaaS	Autonomy
Monitis	SaaS	Autonomy
Rackspace	SaaS	Autonomy
PCMONS	Integrated	Comprehensiveness
CMS	Integrated	Comprehensiveness
mOSAIC	SLA	Accuracy
RMCM	Integrated	Comprehensiveness
MRTG	Basic Metrics	Accuracy
Cacti	Basic Metrics	Accuracy
Nagios	Basic Metrics	Accuracy
FlexACMS	Integrated	Comprehensiveness

2.6 Cluster and Grid Monitoring Solutions

There are clear overlaps between cluster and grid requirements and cloud requirements. For example, clusters, like clouds are composed of many machines connected in local networks. Grids tend to be geographically distributed and belong to autonomous management domains, whereas clouds have a large scale infrastructure managed by a single organization. Cloud specific monitoring solutions have been created to be used in cloud computing environments. Currently, cloud specific monitoring solutions are designed by academic researches [24].

2.6.1 Private Cloud Monitoring Systems:

PCMONS is an open source solution that uses a layer called Integration to provide homogeneous access to users and managers that manipulate resources in a cloud. It provides a uniform monitoring of infrastructure, independently of type of resource hosted in a cloud. In addition, other monitoring solutions can be used as support and complement PCMONS, promoting an integration among monitoring solutions.

2.6.2 Cloud Management System:

CMS aims to provide a monitoring solution based on RESTful Web Services. CMS employs REST to allow the development and integration of monitoring solutions. The REST system can design monitoring elements. The Get method in REST can replace the operations of monitoring [17].

2.6.3 Runtime Model for Cloud Monitoring:

RMCM aims to monitor resources through abstract models, making possible homogeneous handling of heterogeneous resources. However, it requires a constant update of monitoring resources in order to maintain the model consistent. The main disadvantage of this solution is related to the manual installation and configuration of specific agents [18].

3. Conclusion

In this paper, we have presented an overview on cloud monitoring aiming to distinguish the concepts of cloud monitoring requirements and cloud monitoring abilities. Moreover, we presented a comparison among cloud monitoring solutions and discussed trends and future directions in the area to predict a future landscape in order to assist the design and development of new cloud monitoring solutions.

ACKNOWLEDGMENT

Authors are thankful to the managing director Er. T. Ashok Kumar, Principal Dr. Syed Kamaluddin and Dr. K. Ravindranadh, Physics Division, Chirala Engineering College, Chirala for providing research support and financial assistance.

REFERENCES

[1] R. Sobie, Ashok Agarwal, I. Gable, C. Leavett-Brown, M. Paterson, R. Taylor, A. Charbonneau, R. Impey and W. Podiama, “HTC scientific computing in a distributed cloud environment”. In Proceedings of the 4th ACM workshop on Scientific cloud computing, ACM, New York, NY, USA, DOI: 10.1145/2465848.2465850, pp. 45-52, 2013.

[2] Pars Cit: An open-source CRF Reference String Parsing Package. European Language Resources Association, 2008.

[3] H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang and E.A. Fox, "Automatic document metadata extraction using support vector machines: pp. 37-48, 2003.

[4] H. Han, E. Manavoglu, H. Zha, K. Tsioutsiouliklis, C.L. Giles and X. Zhang. "Rule-based word clustering for document metadata extraction". In Proceedings of the 2005 ACM symposium on applied computer, New York, USA, pp. 1049-1057, 2005.

[5] J. Blomer, P. Buncic and T. Fuhrmann. "CernVM-FS: delivering scientific software to globally distributed computing resources". In Proceedings of the first international workshop on Network-aware data management, New York, USA, ACM, pp. 49-56, 2011.

[6] L.M. Vaquero, L. Rodero-Merino, J. Caceres and M. Lindner. "A break in the clouds: towards a cloud definition", SIGCOMM. Comp. Commun. Review, pp. 50-55, 39, 2008.

[7] G. Aceto, A. Botta, W. De Donato and A. Pescape, "Survey cloud monitoring: A survey", Computer Networks, pp. 2093–2115, 57, 2013.

[8] S. Clayman, A. Galis, C. Chapman, G. Toffetti, L. Rodero-Merino, L.M. Vaquero, K. Nagin and B. Rochwerger, "Monitoring Service Clouds in the Future Internet. In Towards the Future Internet - Emerging Trends from European Research", April 2010.

[9] J. Montes, A. Sanchez, B. Memishi, M. S. Perez and G. Antoniou, "Gmone: a complete approach to cloud monitoring", doi:10.1016/j.future.2013.02.011, 2013.

[10] I. Brandic, "Towards self-manageable cloud services", In Proceedings of the 2009, 33rd Annual IEEE International Computer Software and Applications Conference", COMPSAC-2009.

[11] E. Feller, L. Rilling and C. Morin and Snooze, "A scalable and autonomic virtual machine management framework for private clouds", In Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing 2012.

[12] M. Rak, S. Venticinque, T. Mahr, G. Echevarria and G. Esnal. "Cloud application monitoring: The mosaic approach", In Proceedings of the IEEE Third International Conference on Cloud Comp Tech and Sci, 2011.

[13] J. Shao, H. Wei, Q. Wang and H. Mei. "A runtime model based monitoring approach for cloud", In Proceedings of the IEEE 3rd International Conf on Cloud Computing, 2010.

[14] Adrian Cho, Higgs Boson Makes Its Debut After Decades-Long Search, Science, pp. 141-143, 2012.

[15] F.H. Barreiro Megino, D. Benjamin, K. De, I. Gable, V. Hendrix, S. Panitkin, M. Paterson, A. De Silva and R. Walker, "Exploiting Virtualization and Cloud Computing in ATLAS", Journal of Physics: Conference Series, pp. 032011-032018, 396, 2012.

[16] Q. Zhang, L. Cheng and R. Boutaba. "Cloud computing: state-of-the-art and research challenges", Journal of Internet Services and Applications, pp. 7-18, 2010.

[17] J.G. Aguado, "Monpaas: An adaptive monitoring platform as a service for cloud computing infrastructures and services", IEEE Transactions on Services Computing, 2014.

[18] K. Alhamazani, R. Ranjan, K. Mitra, F.A. Rabhi, S.U. Khan, A. Guabtni and V. Bhatnagar, "An overview of the commercial cloud monitoring tools:Research dimensions, design issues and state-of-the-art", CoRR, 1312-6170.

[19] M.B. de Carvalho, R.P. Esteves, G. da Cunha Rodrigues, C. C. Marquezan, L.Z. Granville and L.M.R. Tarouco, "Efficient configuration of monitoring slices for cloud platform administrators", IEEE Symposium, pp. 1-7, 2014.

[20] K. Fatema, V.C. Emeakaroha, P.D. Healy, J.P. Morrison and T. Lynn, "A survey of cloud monitoring tools: Taxonomy, capabilities and objectives", J. Parallel Distrib. Comput., pp. 2918-2933, 2014.

[21] A. Beloglazov and R. Buyya, "Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints", IEEE Transactions on Parallel and Distributed Systems, 2012.

[22] H. Ohman, "Using Puppet to contextualize computing resources for ATLAS analysis on Google Compute Engine", Proceedings of the CHEP Conference, Amsterdam, 2013.

[23] M. Vliet, "Repoman: A Simple RESTful X.509 Virtual Machine Image Repository", Proceedings of the International Symposium on Grids and Clouds, Taipei, 2011.

[24] F. Furano, "Dynamic federations: storage aggregation using open tools and protocols", Journal of Physics: Conference Series doi:10.1088/1742-6596/396/3/032042.

Managing Disaster Event using Geospatial and Web Technologies

G. Rajasekhara Basava Kumar [#1], Nitin Mishra [*], G. Srinivasa Rao [*], V. Bhanumurthy [*], J. V. D. Prasad [#]

[#] *Department of Computer Science, VRSEC*
Vijayawada 520 007, Andhra Pradesh, India

[1] rbkumar516@gmail.com

[*] *Remote Sensing Applications Area, National Remote Sensing Centre-ISRO*
Balanagar, Hyderabad 500 037, Telangana, India

***Abstract*— Disaster Management Portal is a Geographic Information System (GIS) based Web Application to manage Disaster events, which serves the functionalities such as monitoring, tracking, communication, workflow management and utility. It provides the visualization for monitoring different disaster event alerts like Rainfall alerts, Flood alerts, Water level alerts and other relevant alerts. When some disaster event occurs, different activities are associated to manage disaster event in regards Geospatial data handling. This portal will facilitate Activity tracking facility to different concern person to know status of disaster event and also provide satellite tracking facility (track the orbit from different satellites for planning of data acquisition). Communication module can provide communication among the users.**

Keywords: **Disaster Management Portal, GIS, Disaster event alerts, Satellite tracking, Workflow management.**

I. INTRODUCTION

A. Disaster scenario

Natural disasters such as earthquakes, floods, tropical cyclones, wildfire, tsunami and landslides affects different parts of India with varying intensities over space and time. As per the statistics of International Strategy on Disaster Reduction there was an 18 percent rise in disasters during 2005 compared to 2004 [3]. This increase is mainly due to the rising numbers of floods that affect large swathes of population. About 157 million people were affected by disasters in 2005 [3] resulting in damages of about 159 billion USD in the world. India ranks as the second country among disaster prone countries in terms of population affected. India experienced widespread floods, drought, landslides and earthquakes during recent years. Natural disasters are inevitable and it is almost impossible to fully recoup the damage caused by the disasters.

II. DEVELOPMENT OF GIS BASED WEB APPLICATION

Different information technologies are appropriate in the various phases of the disaster management life cycle [1] Imagine a world in which geospatial information is available to all who need it (and who have permission to use it) in a timely fashion, with a user friendly interface [2]. More specifically, technologies should be devised that can help individuals and groups access information, visually explore, analyse and take appropriate decisions.

The development of the web application follows the MVC architecture:

The Model-View-Controller is an architectural design pattern that divides the application into three distinct but interrelated units: the model, the view, and the controller. Each of these components are built to handle specific development features of an application. MVC is one of the most frequently used industry-standard web development framework to create scalable and extensible projects.

A. MVC Components

1) *Model:* This is the business logic of applications, responsible for performing the actual work conducted by the application. Hence, we can say that this unit deals with the modelling of real-world problem and does not have any idea about how it is being displayed to user

2) *View:* This is the presentation logic of application responsible for rendering the information (or data) of the application. It may have little or no programming logic at all.

3) *Controller:* This is the request processing Logic of application, mainly responsible for coupling both the model and view together, so as to perform some operation. We can think it as a traffic controller directing request to the corresponding resources and forwarding appropriate response to the user.

The result of the application design with MVC is the separation of content presentation (View) and content generation (Model) thereby developing maintainable, flexible and extensible application.

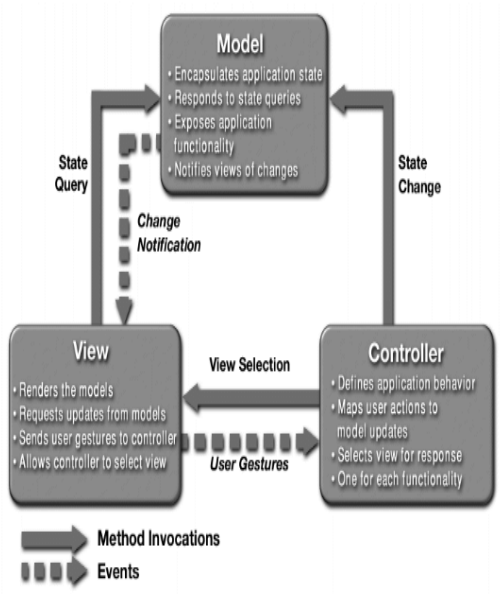


Fig 1: MVC Architecture [4]

The following web technologies and open source softwares will be the some technologies to make the Disaster Management Portal.

TABLE I
DIFFERENT TECHNOLOGIES USED IN APPLICATION DEVELOPMENT

Component	**Open source software (Freeware)**
Web Server	Glassfish/Tomcat
Database	PostgreSQL
Spatial Data Base Engine	PostGIS
Server Spatial Data Publishing Engine	GeoServer
Server Side Development Environment	JSP, Java Bean.
Client Side Development Environment	HTML5, JavaScript, jQuery, AJAX, CSS3, Leaflet
Data Structure for Spatial Data Communication	GeoJSON

The framework made for the Disaster management portal having the modules like Disaster alerts, Satellite tracker, Workflow mapper, Event Inbox, Event archival, Activity tracking, Utilities and Help module.

The web application has the database, client side environment and the server side environment. Hence the development of web application starts with the database creation.

III. Database Creation

Much of the information that is required for emergency preparedness, response, recovery, and mitigation including resources allocation involve geospatial information. Different information technologies are appropriate in the various phases of the disaster management life cycle [6] Imagine a world in which geospatial information is available to all who need it (and who have permission to use it) in a timely fashion, with a user friendly interface [2]. More specifically, technologies should be devised that can help individuals and groups access information, visually explore, analyse and take appropriate decisions.

Databases are designed that allow for the storage and retrieval of large quantities of related data. Databases consist of tables that contain data. When creating a database one should think about what tables that going to create and what relationships exist between the data in the tables. A good database design will ensure the integrity and maintainability of data that is stored in the database.

Relational databases are designed for fast storage and retrieval of large amount of data.

Some of the features of relational databases and the relational model are following:

- Use of keys
- Constraining the input
- Maintaining data integrity
- Structured Query Language(SQL)

IV. Web Application Features

A. *Web Mapping*

Web mapping is the process of using maps delivered by geographical information systems (GIS). A web map on the World Wide Web is both served and consumed, thus web mapping is more than just web cartography, it is a service by which consumers may choose what the map will show. Web GIS emphasizes geo-data processing aspects more involved with design aspects such as data acquisition and server software architecture such as data storage and algorithms, than it does the end-user reports themselves. The terms web GIS and web mapping remain somewhat synonymous. Web GIS uses web maps, and end users who are web mapping are gaining analytical capabilities. The term location-based services refers to web mapping consumer goods and services. Web mapping usually involves a web browser or other user agent capable of client-server interactions.

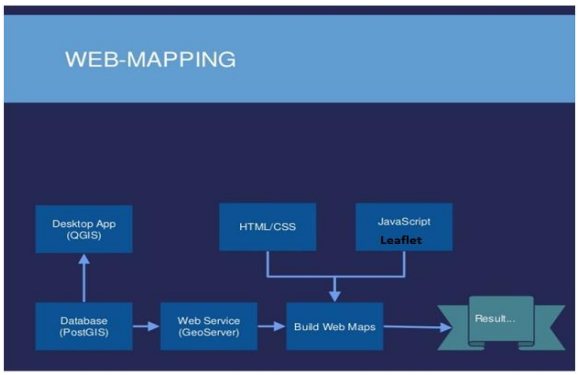


Fig: Web Mapping

B. Displaying Map Data in Browsers

The map/spatial data should be displayed on the web browsers using the JavaScript libraries or the plugins. Some of the examples of the JavaScript libraries or the plugins for map displaying in the browsers including the Leaflet, OpenLayers, MapBox, MapQuest Maps etc.,

The map library API's providing us many features to work on the maps. Those features makes easy to work on the geospatial things. These features includes the geotagging, drawing vector features on the map, using of different WMS/WMTS services, loading the shape files or KML files or GeoJSON files.

A. Drawing Vector Features on the Map

There is facilitation for drawing the vector layers in the map to show the disaster event locations. The vector layers includes the drawing of polygon, polyline, circle, marker, rectangle etc.,

The features drawing on the map are sent to the database and stores in the database in the geometry format. The additional facilities including the editing of the vector layers and deleting the drawn vector layers before sent into the database.

B. Providing the Geotagging facility

One of the most important facilities on the geospatial map is the geotagging. It is the process of adding geographical information to various media in the form of metadata. The data usually consists of coordinates like latitude and longitude, but may even include bearing, altitude, distance and place names.

In the disaster management portal, we can add the information related to the disaster event on the event location and we can show on the map using the popup.

C. Population of map with the GIS data

The map in the web browser is populated with the GIS data that is stored in the spatial database. The populated data consist of the geometry data, GeoJSON data format, text data format and may be the image data format. The spatial data related to the disaster events can be populated over the map, then it will be easy to know about the disaster event locations and the working on the disaster event can also become easy.

D. Providing Communication among the users

The users who are working on the disasters, the communication is very important. There are lot of actions to be performed during the disaster management. The communication can be achieved through the SMS, mail and through the communication module in the Disaster Management Portal. The users can be divided into Project coordinator, Project manager, Special users and the Guest users. The communication among the users is based on the event. Some specific people are for some events, they can work on their corresponding assigned events.

E. Providing different privileges to the different users

There are different privileges for the different users because the Disaster Management Portal can be different kinds. Different activities are assigned to the different users. Based on their activities there is differences in providing the privileges to the users.

First there is a work flow for the disaster management portal, based on the workflow project can be defined.

1) Project Coordinator:

The following will be the privileges for the Project coordinator:

a. Project coordinator will create the event based on the disaster alerts and the news/ground information.
b. In the creation of event project coordinator will see the disaster alerts and track the satellite information with the help of the satellite tracker.
c. Based on the disaster event, project coordinator will assign the project manager and the event members.
d. Project coordinator can selects the event based satellites.
e. Project coordinator have the facility to draw vector layers over the map and also have the Geotagging facility to tag necessary information related to the disaster event location.
f. Whatever the event description and the event source given by the project coordinator that will be displayed on the Recent News module.
g. Project coordinator can have the facility to assign the new project manager to the created event, whenever the assigned project manager has rejected the event.
h. Project coordinator will communicate with the project manager and event members with the help of the messaging.
i. After the completion of the project, Project coordinator will close the project.
j. Report generation is the final stage of the project. Project coordinator generates the report related to the event based on the timeline activities that has done during the disaster activation time.

2) Project manager:

The following are the privileges defined for project manager:

a. Project manager having the facility to see the satellite tracker and plan according to the satellite pass.

b. Project manager will accept/reject the event when project coordinator has assigned. Project manager can write the reason for rejecting the event.
c. Project manager works on the disaster event and updates the status of work based on the event.
d. The communication is among the event members and the project coordinator can be achieved based on the event.
e. Project manager updates the status on the activity tracking module which can defines the timeline activities of the project.

3) Special users:

The following are the privileges defined for special users:

a. Special users will monitor the disaster event from the creation of event to the end of the event.
b. They can also communicate with the event members with the communication module.
c. They can also observe the disaster alerts and satellite tracker, then the communication among event members can also be possible.

4) Guest users:

Guest users will have some specified login privileges. Guest users can have the facility to observe the disaster management activities on the portal and they are having limited privileges compared to the other users who are working on the disaster events.

V. Disaster Management Application Architecture

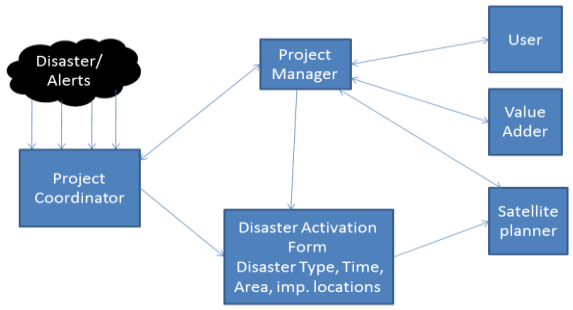


Fig: Disaster Management Portal Architecture

The workflow of the disaster management application:

a. Project Coordinator Activates the process
   - Based on alerts provided by the system
   - Based on news/ground information
b. Project Coordinator identifies Project Manager
c. Software automatically generates
   - Event-ID
   - Possible satellite coverages (visualization, table)
   - This info is accessible to Project Coordinator, Project Manager, Satellite Planner
d. Satellite Planner finalizes the planning in consultation with Project Manager for pre & post data.
e. Satellite Planner supplies pre disaster data immediately through FTP.
f. Project Manager prepares historic and forecasting scenarios and disseminates to user
g. Informs the user regarding planned products. Receives user requirements and accordingly makes plans for generating products
h. Satellite planner supplies post disaster data through FTP
i. Project manager prepares value added products and disseminates to user
j. Repeat steps-7-9, till the project manager recommends closure of the activation in consultation with user and sends the final report to project coordinator
k. Project coordinator closes the activation.

VI. Conclusion

This paper mainly focused on managing the disaster events using the geospatial and web technologies. The portal facilitates the different privileges to different users based on their role. Geospatial model can help to identify the locations and information related to the disaster event. This application can help the users for tracking the disaster event to plan their further activities. Different modules in the application provides different services and facilities which are designed in user friendly way, that makes easy in using of the application/portal. The report is generated at the final stage of the project based on the timeline activities involved in the project.

Acknowledgment

The authors express their sincere thanks to Scientists, Remote sensing Applications Area, NRSC-ISRO and very much thankful to Management and principal of VR Siddhartha Engineering College, Vijayawada for their support and constant encouragement.

References

[1] Committee on Using Information Technology to Enhance Disaster Management, National Research Council (2005) Summary of a Workshop on Using Information Technology to Enhance Disaster Management, National Academies Press. http://www.nap.edu/catalog/11458.html
[2] Committee on Intersections Between Geospatial Information and Information Technology, National Research Council (2003) IT Roadmap to a Geospatial Future, National Academies Press. http://www.nap.edu/catalog/10661.html
[3] www.unisdr.org
[4] http://java.sun.com/blueprints/patterns/MVC-detailed.html
[5] V. Bhanumurthy , G Srinivasa Rao, "EMERGENCY MANAGEMENT – A GEOSPATIAL APPROACH", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B4. Beijing 2008

Configuration Monitoring and Auditing Of LAN Switches

G. Krishna Kishore [#1], D.S Lakshmi [*2], M.Narasimha [#3]

CSE-DCCT ITD, VRSEC-DRDL
kanuru Vijayawada India-Kanchanbagh Hyderabad India
[1]gkk@vrsiddhartha.ac.in
[2]slakshmidivakaruni@gmail.com
[3]mnarasimha@drdl.drdo.in

***Abstract*— DRDL campus wide Gigabit fibre optic LAN consists of 180 switches of make Cisco and D-link combine and 1500 nodes. All these switches of type Layer2, Layer3 & containing various releases of Cisco IOS. Day to day activity of network maintenance engineers has to configure and maintain secure features as per lab's information security policy. To ensure this, Network management periodically checks & audits all these switches. It takes lot of time to complete this operation. Hence, to ease this activity, it is required to develop a web based software tool to configure, monitor and report. The main issue of auditor is logging in to 180 switches with ip address & entering logging credentials using SSH software and verifying secure features configurations and recording manually consumes a lot of Auditor time. These secure features that need to verify are ip address, VLANs, MAC address, port connectivity status, port modes (access, trunk) and disable of unwanted services like telnet, http, etc. These parameters are to be recorded in a database daily. Using this database changes report is to be generated and reasons for the changes to be recorded. The tool will be developed using HTML, Java servlets, Java scripts, and Oracle Database.**

Keywords: SSH, SSH History, SSH Configuration, SSH working, Java secure channels, putty configuration tool.

I. INTRODUCTION

Switches are an Ethernet based LAN [11] devices, it reads incoming TCP/IP data packets/frames containing destination information as they pass into one or more input ports [1]. The network engineer's work is to detect the problems occurred in DRDL campus wide switches.

The main issue occurred in this process was, the network engineers manually must go to that location to solve the problem which consumes more time. Hence, to ease this process, there is a need to configure the switch using various programs exist for remotely login purposes, such as telnet, rlogin. As many of these network-related issues have major problems, they lack security. If you logged in to other computer remotely using telnet, your username and password can be known easily as they travel over the internet.

SSH [2] is a powerful, popular, software-based approach to network-security. Whenever a data is sent through computer to network, SSH automatically encrypts it. When the data reaches its known recipient, SSH automatically decrypts [2] (unscrambles) it. The result is transparent encryption: In addition, SSH uses modern, secure encryption algorithms and is effective enough to be found within mission-critical applications at major organizations like DRDL.

SSH uses client-server architecture, as shown in fig. 1[3] An SSH server is typically installed and can be run by system administrator, accepts, or rejects incoming connections to its host computer. Users then run SSH client programs, typically on other computers, to make requests of the SSH server.

All communications between clients and servers [3] are securely encrypted and protected from modification. SSH clients communicate with SSH servers over encrypted network connections.

The use of SSH [9] in this project is to help the network engineers to get data of all switches in a secured manner. All the information in SSH protocol will transfer in encryption format, hence third party cannot be able to hack the secured information.

The information in SSH configured switch helps network engineers to audit the desired results that obtain from 180 switches in to a browser designed. These information helps them to bind the Mac address, to configure vlan number for switch ip address, to know status of a port modes (access, trunk), whether the port is shutdown or not.

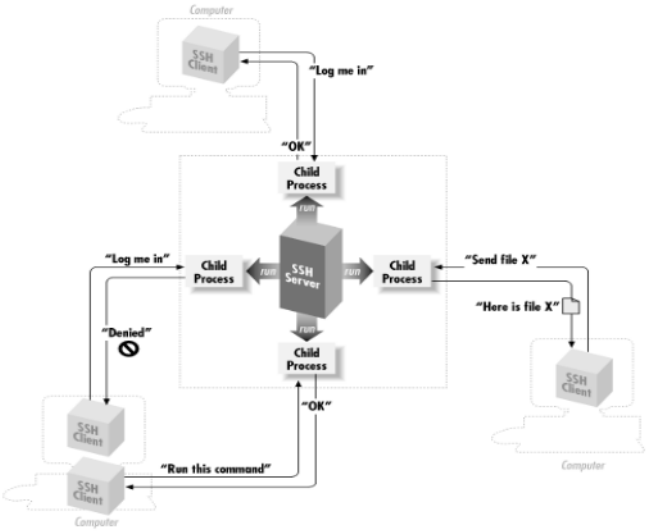


Fig. 1. SSH Architecture

The fig. 2 [3] helps us to know how SSH provides authentication, encryption, integrity.

A. Authentication

Reliably determines someone's identity. If you try to log into an account on a remote computer, SSH asks for digital proof of your identity. If you pass the test, you may log in; otherwise SSH rejects the connection.

B. Encryption

Scrambles data is unintelligible except to the intended recipients. This protects your data as it passes over the network.

C. Integrity

Guarantees the data travelling over the network arrives unaltered. If a third party captures and modifies your data in transit, SSH detect this fact.

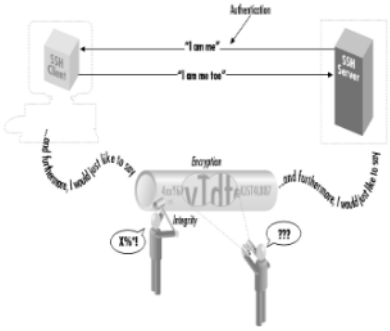


Fig. 2. Authentication, Encryption and Integrity

II. RELATED WORK

SSH1 and the SSH-1 protocol were developed in 1995 by Tatu Ylonen [18], a researcher at Helsinki University of Tech. in Finland. After this his university network was the victim of a password-sniffing attack earlier that year. He realized that his security product could be put to wider use.

As in this same year he documented the SSH1 protocol as an Internet Engineering Task Force (IETE) [18] Internet Draft, which essentially described the operation of SSH1 software. It leads to several limitations, so in 1996, SCS introduced a new version of the protocol that is SSH 2.0 or SSH-2. Which incorporates new algorithms and is compatible to SSH-1.

In 1998, SCS released the software product "SSH secure shell" (SSH2). SSH2 did not replace with SSH1 i, for 2 reasons. First SSH2 was missing a number of useful, practical features and configuration options of SSH1. Second, SSH2 had a more restrictive license. SSH2 is a better and more secure protocol.

OpenSSH [19] is gaining prominence as an SSH implementation, developed under OpenBSD [20] project and freely available under license. It supports both SSH-1 and SSH-2 in a single set of software. OpenSSH has been ported successfully to Linux, Solaris, AIX, and other operating systems, in tight synchronization with main releases.

Related technologies that were used before SSH are rsh Suite (R-Commands) [3], pretty good privacy (PGP) [3], Kerberos [3], IPSEC [3], Secure Remote Password (SRP) [3], Secure Socket Layer (SSL) Protocol [3], SSL-Enhanced Telnet and FTP [3], Stunnel, firewalls.

SSH [8] provides more security for remote connections than telnet does by providing strong encryption when a device is authenticated. This software releases support SSH version1 (SSHv1) [4] and SSH version2 (SSHv2) [4]. The SSH feature has an SSH server and SSH integrated client, which are applications that run on the switch.

The SSH [12] server works with the SSH client supported in the release and with non Cisco SSH clients. The SSH client also works with the SSH server supported in this release and with non-Cisco SSH servers [4]. The switch supports an SSHv1 or an SSHv2 server. The switch supports an SSHv1 client [4]. SSH supports the data encryption algorithm and password-based user authentication. SSH also supports local authentication and authorization methods [4].

How SSH works? [15], [13] When you connect through SSH [8], you log in using an account that exists on the remote server. When you connect through SSH, you will be dropped into a shell session, which is a text-based interface [5] where you can interact with your server. For the duration of your SSH session, any commands that you type into your local terminal are sent through an encrypted SSH tunnel [12] and executed on your server.

The SSH connection [5] is implemented using a client-server model. This means that for an SSH connection to be established, the remote machine must be running a piece of software called an SSH daemon [5]. This software listens for connections on a network port, authenticates connection requests, and spawns the appropriate environment if the user provides the correct credentials [15].

The user's computer must have an SSH client. This is a piece of software that knows how to communicate using the SSH protocol [5] and can be given information about the remote host to connect to, the username to use, and the credentials that should be passed to authenticate. The client can also specify certain details about the connection type they would like to establish [13].

How SSH Authenticates Users? [5] Clients generally authenticate either using passwords (less secure and not recommended) or SSH keys, which are very secure. Password logins are encrypted and are easy to understand for new users. However, automated bots and malicious users will often repeatedly try to authenticate to accounts that allow password based logins, which can lead to security compromises. For this reason, we recommend always setting up SSH-based authentication for most configurations.

SSH [9] keys are a matching set of cryptographic keys which can be used for authentication. Each set contains a public and private key. The public key can be shared freely without concern, while the private key must be vigilantly guarded [5] and never exposed to anyone. To authenticate using SSH keys, a user must have an SSH key pair [9] on their local computer. On the remote server, the public key must be

copied to a file within the user's home directory. This file contains a list of public keys, one-per-line, that are authorized to log into this account.

When a client connects to the host, wishing to use SSH key authentication [5], it will inform the server of this intent and will tell the server which public key to use. The server then checks its authorized_keys [5] file for the public key generate a random string and encrypt it using the public key. This encrypted message can only be decrypted with the associated private key.

The server will send this encrypted message to the client to test whether they actually have the associated private key. Upon receipt of this message, the client will decrypt it using the private key and combine the random string that is revealed with a previously negotiated session ID. It then generates an MD5 [6] hash of this value and transmits it back to the server. The server already had the original message and the session ID, so it can compare an MD5 [6] hash generated by those values and determine that the client must have the private key.

Guidelines to configure a switch as an SSH server or SSH client: An RSA key pair [4] generated by SSHv1 server can be used by SSHv2 server, and the reverse. If you get CLI error message after entering the crypto key generate RSA global configuration command, an RSA key pair has not been generated. Reconfigure the hostname and domain, and then enter the crypto key generate RSA command.

When generating the RSA key pair [4], the message no host name specified might appear. If it does, you must configure a hostname by using the hostname global configuration command. When generating the RSA key pair, the message domain specified might appear. If it does, you must configure an IP domain name by using the ip domain-name global configuration command. When configuring the local authentication and authorization authentication method, make sure that AAA [7] is disabled on the console.

Steps to configure SSH switch: This procedure is required if you are configuring the switch as an SSJ server.
Configure terminal is used to enter global configuration mode.
Hostname is used to configure a hostname for your switch.
Ip domain-name is used to configure a host domain for switch.
Crypto key generate rsa used to enable the SSH server for local and remote authentication on the switch and generate an RSA key pair.

We recommended that a minimum module size of 1024 bits. End, it returns to privileged EXEC mode. Show ip ssh or show ssh: this command describes about the version and config information for your SSH server. Show SSH describes the status of the SSH server connections of the switch. Copy running-config startup-config: (optional) save your entries in the configuration file.

Steps to configure the SSH server: Configure terminal, ip ssh version [1 2]: configure the switch to run SSH1 or SSH2.if you do not enter this command or do not specify a keyword, the SSH server selects the latest SSH version supported by the SSH client. For Example, if the SSH client supports SSHv1 and SSHv2, the server selects SSHv2.

Ip ssh {timeout seconds | authentication-retries number}: Specify the timeout value in seconds, the default timeout is 120 seconds the range is between 0 to 120 seconds. Line vty line_number [ending_line_number]: configure the virtual terminal line number. Enter line configuration mode to configure the virtual terminal line settings. For line_number and ending_line_number, specify a pair of lines. The range is 0 to 15. Transport input ssh: specify that the switch prevent non-SSH telnet connections. This limit the router to only SSH connections. End, return to privileged EXEC mode. Show ip ssh or show ssh, copy running-config startup-config.

Displaying the SSH configuration and status: Show ip ssh is used to show the version and configuration information for your SSH server. Show ssh: show ssh command shows the status of the SSH server connections on the switch.

III. IMPLEMENTATION

Java secure channels (JSch) [16] are a new concept used to get SSH configured switch details. It is a pure java implementation of ssh2. The jsch-ssh2 plug-in for Grails provides a basic SSH client using the JSch library provided by Jcraft [14] jsch-ssh2 provides an easy to use client for running commands on a remote host, and copying files to and from a remote host.

This plug-in strives to provide an easy convention to follow quickly begin running commands [14] on remote hosts, sending files to remote hosts, and fetching files from remote host using the SSH protocol. It is a free SSH client library for the java environment. JSch [16] supports multiple channels (operations) over a connection to the server. JSch allows you to connect to an sshd server and use port forwarding etc., and you can integrate its functionality into your own java programs.

There are some of the channel methods that helps to implement SSH using JSch those are user interface, PromptYesNo, Promptpassword, setPassword getUsername, sessions and config commands, config host key, open channel, input channel, output channel, session timeout, print stream, port forwarding: local, remote and dynamic, exec helps to execute a command asynchronously on this channel. Request_pty is used to request the terminal be opened for a channel.

Send_data is used to send the given data string to the server via channel. Subsystem is used to request the server to start the given subsystem on this channel. Send_request is used to send a named request to the server for a channel. Send_signal indicates that the server should send the given signal to the process on the other end of the channel.

Send_extend_data is used to send a data string to the server, along with an integer describing its type. Send_eof is used to tell the server that no further data will be sent from the client to the server. Exec is used to execute a command asynchronously on this channel.

The methodology that is used to implement this project was ip address, username, password, enable password. All these details are passed to JSch sessions and channels in order to avoid login credentials for each SSH configured switch.

Among these details ip address, interface, vlan number, Mac address, shutdown and current date features are needed for network engineers to monitor.

Based on current date feature helps engineers to know status of switch information of previous date. Here Day by Day information will not be update but it will be stored in the database for future use and this information will display on the web based software tool to monitor, configure and report.

The network engineer's uses putty [10] software as shown in Fig. 3 to log into a switch using an ipaddress. After login to the putty software [10] using ipaddress of a switch, then click the SSH button and click on open button in order to login in to switch.

This redirects to enter switch ip credentials such as username, password. If the user is authorized, then it enters into user mode (>) else displays the access denied message. Now enter the enable password to enter into privileged mode. Enter the second password, by this password we will enter into the switch. Next, enter the show run command, which is used to know the building configuration of the particular switch.

The building configuration displayed in the switch mainly consists of rsa certificate which is used to provide the security to the Ssh. Using this configuration we can know that which port in the switch is connected to the computer. Next, enter the exit command, which helps to come out of the switch.

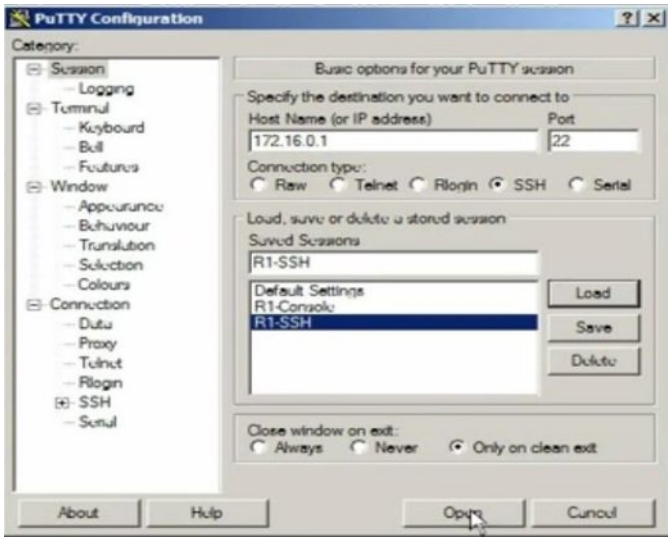


Fig. 3. Putty configuration tool

Implementation of this project based on four modules those are input, array, database and reports. In input module, we gather 180 switches ip & credentials from network engineers and keep this information in to a file. Read this file and store in to a buffer. Now read data from buffer and split those values. Now pass these splitted values to the java secure channel (sessions) and to the print stream function in order to get switches configuration details.

In array module, retrieve the switch details from file and read this file using buffer reader. Store this file into a variable, in such a way to split that variable for future use. If that variable contains ipaddress, print that value and add to java object, now repeat the same process for interface, vlan, mode, and Mac address & shutdown parameters and add these java objects to the array object in order to insert these parameters into database.

In database module, we establish a database using jdbc. By using database connections we get ipaddress, interface, vlan, mode, mac address, shutdown parameters. Now generate an insertion query, which includes all these parameters.

In reports module, generate the reports according to requirements those are All switches current configuration details, particular switch configuration details, particular vlan configuration details, total mac address, list of ports not secure.

IV. RESULTS

Fig.4. Input page

In this page we can view run button and reports link. If we click on Run button it process the current switch details running in the background. When we click on Reports link it displays all switches information.

Fig.5. Selecting date page.

In this page, if we choose any particular date then it displays the desired date switch information such as ipaddress, interface, vlan, mode, Mac_address, shutdown and date.

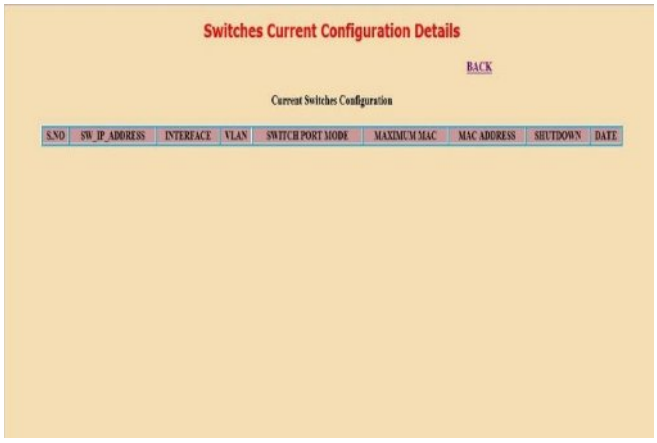


Fig.6. All switches current configuration details

In this page, It displays all switches current configuration details such as, ipaddress, interface, vlan, mode, Mac_address, shutdown and date.

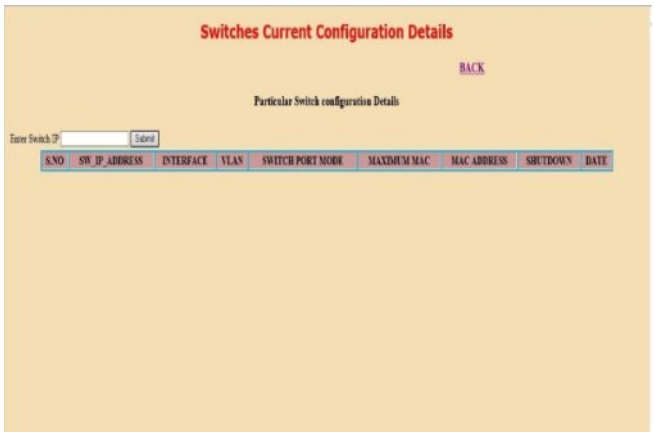


Fig.7. Particular switch configuration details

In this page, when we enter the paricular swicth ipaddress it displays the details of switch such as, ip, interface, vlan, mode, Mac address, shutdown and date.

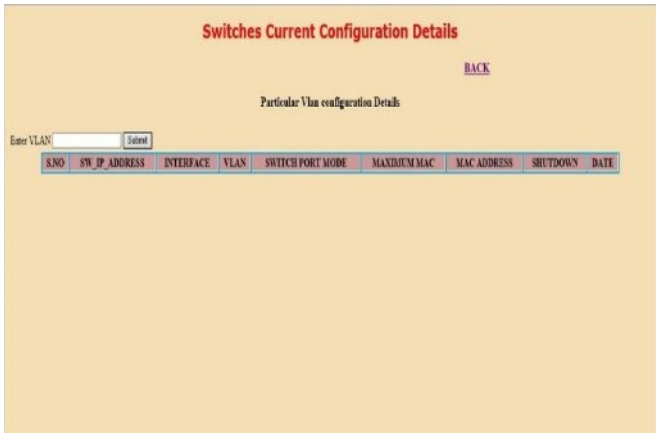


Fig.7. Particular vlan configuration details

In this page, when we enter the particular vlan number it displays the details of that valn. The details to be displayed are ipaddress, interface, vlan, mode, Mac address, shutdown and date.

Fig.8. Total Mac_Address

In this page, it displays mac addresses of all the switches such as, ipaddress, interface, vlan, mode, mac address, shutdown and date.

Fig.9. List of ports not secure

In this page, it displays all switches information if that switches connected in access mode. There are two modes, access and trunk. Access represnts a switch port connected to pc. Trunk mode represents a switch is directly connected to another switch. Both these switches uses fiber optic cables or gigabit ethernet cables.

V. CONCLUSION

With the system discussed in this paper, Configuration monitoring and auditing of all the switches is an achievable and successfully utilizing in the organization. By this project network engineers are auditing and monitoring of all switches. They are easily identifying whether particular switch is up or down. Whether the switch contains all ports secured or not. If any of those switches are not secured then mac binding should be done for that port. From above results, List of ports not

secure page shows, whether the switch is in access or trunk mode. Total mac address page shows the mac addresses of all switches. All switches current configuration page shows the details all the switches. Particular switch, vlan configuration pages shows switch ip details and vlan number of a particular switch that given.

REFERENCES

[1] Sean Wilkins “An introduction to network devices” Oct 2015.
[2] T. Ylonen and C. Lonvick, “The Secure Shell (SSH) Protocol Architecture”, RFC 4251, Jan. 2006.
[3] O’Reilly “SSH: The Secure Shell”, ISBN 0-596-00011-1 Feb 2002.
[4] ConfigSSH,http://www.cisco.com/c/en/us/td/docs/switches/lan/catalyst2960/software/release/12-240se/configuration/guide/scg.pdf#G11.1227177.
[5] SSHServers,http://www.digitalocean.com/community/tutorials/ssh-essectials-working-with-ssh-servers-clients-and-keys.
[6] R.Rivest “The MD5 Message Digest Algorithm” RFC-1321 April 1992.
[7] AAAModel,http://www.cisco.com/c/en/us/td/docs/ios/12_2/security/configuration/guide/fsecur_c/scfssh.html
[8] Oliver Gasser and Ralph Holz “A Deeper Understanding Of SSH: Results From Internet Wide Scans” ISBN 978-1-4799-0913-1 June 2014.
[9] Polydistortion.net “A Basic Introduction to SSH” April 2014
[10] Putty, http://www.chairk.greenend.org.uk/~sgtatham/putty/.
[11] LANs,http://www.lantronix.com/resorces/networking-tutorials/ethernet-tutorial-networking-basics/.
[12] Kimmo Suominen “Getting Started with SSH” July 2004.
[13] SSH,http://www.cisco.com/c/en/us/support/docs/security-vpn/secure-shell-ssh/4145-ssh.html.
[14] JSch,http://epaul.github.io/jsch-documentation/simple.javadoc/index.html?com/jcraft/jsch/UserInfo.html.
[15] SSHWork,http://www.slashroot.in/secure-shell-how-does-ssh-work.
[16] JSch,http://www.jcraft.com/jsch”Introduction to Jsch.
[17] JSch,http://www.itmagix.blogspot.in/2015/05/java-secure-channel-introduction-jsch.html.
[18] O’Reilly “SSH: History of SSH”, ISBN 0-596-00011-1 Feb 2002.
[19] OpenSSH, http://www.openssh.com/ Dec 1,1999.
[20] OpenBSD, http://www.openbsd.org/ May 4,2000.

Impact of Location Popularity on Throughput and Delay in Mobile Ad Hoc Networks

Dr. G. Krishna Kishore[1], R. Navya[2], Rajesh K[3]

[1]*Associate Professor, Dept of CSE, VR Siddhartha Engineering College*, Email: gkk@vrsiddhartha.ac.in

[2]*PG Scholar, Dept of CSE, VR Siddhartha Engineering College*, Email: navyarai793@gmail.com

[3]*Assistant Professor, Dept of CSE, DVR & Dr. HS MIC College of Technology*, Email: rajeshkalakoti@gmail.com

***Abstract*—Now-a-days every smart device is based onlocation. The mobile users are dependent on different locations. users are most likely to visit popular locations. According to the location based scenario we calculate throughput and delay under multi-hop where before studies shown 2-hop got negative performance and 3-hop slightly decreased the delay. So to increase throughput and decrease delay the multi-hop relay algorithm is used. By that the throughput and delay calculations can say that network performance is increased.**

Keywords: Mobile Ad Hoc Networks (MANETs), Location Popularity, Throughput and Delay.

I. INTRODUCTION

Now-a-days smart portable devices are becoming more popular. The Process of communication is the information store and then sending.it is delay. To transfer by that it is an delay so delay-tolerant networking [1]. This DTN has many applications including but it not limited to vehicular networks [2]. And also a pocket switched networks [3], and Sensor networks [4] and rural kiosks which means very small stores and it provides internet access in the countries which are developed [5].

Despite the disruptive nature and on-and-off connectivity of MANETs, it is shown that a constant per-node throughput can still be achieved by exploiting using mobility according to Gross glauser and Tse's work [6]. Since then, there is a huge interest in studying how to use the mobility properties to improve network communications. Performance of network is decided by the information delivery concept and also which is strongly and highly related to the mobility patterns.

Different mobility models are there ranging from simple independently model, identically distributed model and also more complex random mobility models, such as the Brownian motion model, random way-point ad versions of random walk. Neely and modiano study delay-capacity trade-offs in a cell-partitioned structure under i.i.d mobility model.

They develop a scheduling scheme using unnecessary packet transmissions to reduce delay at the expense of capacity. A necessary trade-off between delay and capacity is established, i.e., delay/capacity.

In before works through the entire network area considered nodes are moving uniformly. Assumptions are not hold by the practical settings. In understanding the mobility patterns role in wireless communications the prior studies made great contributions and also some additional dimensions are needed to exploit. Example: location heterogeneity.

Garetto investigate the impact of restricted mobility on network performance. each and every node moving around the home point where it becomes large when occurrence probability decays as the distance from the home point.

For every particular node, the different locations representing different occurrence probabilities after that this information can also be used to construct the scheduling schemes. So this model have some sort of location heterogeneity. In the whole network, the node distribution still having uniform over all locations.

So the four square is the location based application it becomes more popular. So number of traces which it records the particular user visiting. To study the user mobility patterns this traces are retrieved by many researches. observing the prior studies shown the some locations are visited frequently and some are less.

For example, there are number of students in class-rooms, auditoriums and libraries than in the street. So each user have their own different location to visit. It is their popular visit. The popularity indicates a different mobility pattern. And it leads to non-uniform node distribution. This significant difference from previous works suggests that conventional schemes of the MANETs may not be then directly applied to the performance analysis in MANETs with location popularity.

II. RELATED WORK

Capacity and delay tradeoffs for ad hoc networks [7] paper have to achieve capacity of the network an algorithm is developed. This algorithm has a restriction on the number of hops in the network. Only two hop paths are allowed for packet transmission.

In the first hop any users who are available in the network can receive packets. The remaining packets that are stored in buffer are transferred only when they get the opportunity to reach destination.

Cell partitioned relay algorithm explains, for each cell there exists at least two users. The packet are transferred is the newly arrived packet is intended for destination. If the destination is unknown then select any user as destination in the network. Here it has an advantage capacity for two hop relay is achieved by taking concepts in to consideration. It has disadvantage of queuing information is always required.

Mobility increases the capacity of ad hoc wireless networks [8] paper has an algorithm to improve throughput for nodes with mobility other than fixed nodes. Relay concept of packets is introduced to achieve throughput. Packets are transferred to destination through different relay nodes. In case of fixed nodes choosing of destination is random in the network.

In order to increase the throughput direct communication is not sufficient so relaying should be done. It has advantages of maximum throughput can be achieved two hop paths. Traffic spreading is done in order to relay the packets to destination. It has a disadvantages of delay factor is not unknown then the packets must be compulsory relayed to destination.

Capacity and delay in mobile ad hoc networks with f-cast relay algorithms [9] paper has two hop relay algorithm is proposed in which packet is delivered to f-distinct relay nodes.

This the algorithm contains f+1copies of packet exist at each node. Here f+1 copies include source node copy. Every node contains n number of queues. One queue is locally generated and waits for redundant copies and relay queue that contain n-2 relay nodes.

Packets are transferred from source to destination if destination is one-hop neighbour of source or else relay nodes are selected. Here sending number and receiving number of packets are considered in order to get original message from source. It has advantages that Improves network throughput.

Maximum per node throughput can be determined by the corresponding optimal setting of f. It has disadvantages of Packets are delivered to f distinct relay nodes, so it increases time. Congestion problem occurs due to redundancy.

Delivery probability of two hop relay with erasure coding [10] paper has two hop relay algorithm for achieving delivery probability as 0 or 1. 0 indicates the packets are not delivered and 1 indicates that the packets are delivered successfully. If destination is within the range then source transfer the packets to destination directly or else a random relay node is selected. Given messages is divided into same sized frames.

A model markov chain is developed for transferring frames. Markov chain frame work contains four cases. Source to destination, source to relay, relay to destination, source to relay and relay to destination.

It has advantages of It reduces delay of the message. Increases delivery ratio by minimizing the message lifetime. Increases throughput capacity. Markov chain frame work has disadvantages of that if the selected relay node is not the correct node for transmitting frames to destination then it can't take the advantage of each transmission. Selection of relay node is not specified.

Message drop and scheduling in DTNs [11] Algorithm for dropping of the messages and scheduling of the messages in delay tolerant networks is proposed. To decide the message order scheduling is required. When queue is full it is important thing to decide about the packets that are dropping.

III. METHODOLOGY

3.1 Network Model

A. Cell-partitioned Network Model

Considering a cell partitioned network model as in fig 1. Where there is n number of mobile nodes in unit square. The entire network which of equal area is divided into the n non-overlapping cells. Every cell assigned with popularity. Every node can visit the aligned cell by its popularity and it is different non-moving distribution. Nodes which are in the same cell only can communicate with each other node. For every cell the per timeslot and only on transmission is restricted. Different frequencies are there among that neighbouring cell to avoid interference. For the network the only four frequencies are enough.

B. Mobility Model

There are so many mobility models. Here the location-popularity model is used. Time divided into equal duration. After each timeslot, the position of the nodes is totally reshuffled, independently from slot to slot and it is among nodes. According to cell's popularity the node moves to a new cell at the beginning of each timeslot. And for the entire slot duration it remains in the cell. The nodes carries packets until it reaches destination with the help of mobility.

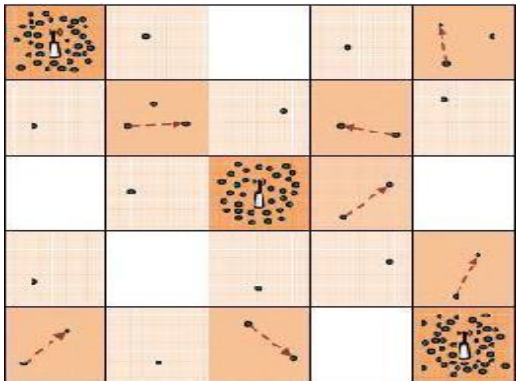


Fig 1: A cell-partitioned MANET model

3.2 Proposed Work

Investigating the throughput and delay using multi-hop relay algorithm based on location popularity and using access points to control the traffic in popular cells for stability of network. By using this algorithm calculating the packet delivery ratio packet loss and routing overhead.

Throughput:

The average number of packets that transferred from every node to destination per unit time is called as throughput. The sum of all per node throughput over all the nodes in a organization is called as the throughput of the network.

Average Packet Delay:

Time taken by packet to reach its destination after it needle the source. The average packet delay of a network is access by averaging over all broadcast packets in the network.

Packet Delivery Ratio (PDR):

It is the ratio of number of packets reached successfully at destination to the number of packets sent by source node.

Data Transmission Energy:

It is the energy of node appropriate to carry data from source to its destination. The average data communication energy of a network is achieved by averaging over all energy in the network.

IV. IMPLEMENTATION

Network Model:

NS2 simulator is used for implementation. Considered number of nodes as 17. These nodes are moving randomly. In the network every node have x and y parameters so by that can identify the location of each node. Considered 17 nodes are numbered as from 0 to 16 and indicated by circle. And here considering node s is selected as source and d as destination. Remaining are intermediate nodes in fig 2 it shows clearly. For simulation taken parameters are given below table

Queue Capacity	50 Packets
Packet Size	1000kB
Packet Interval	0.00001 Sec
Initial Energy of Node	100J
Transmission Power	0.2J
Sleep Power	0.001J
Simulation Area	300x300
Simulation Time	120 Secs

Table 1 Simulation Parameters

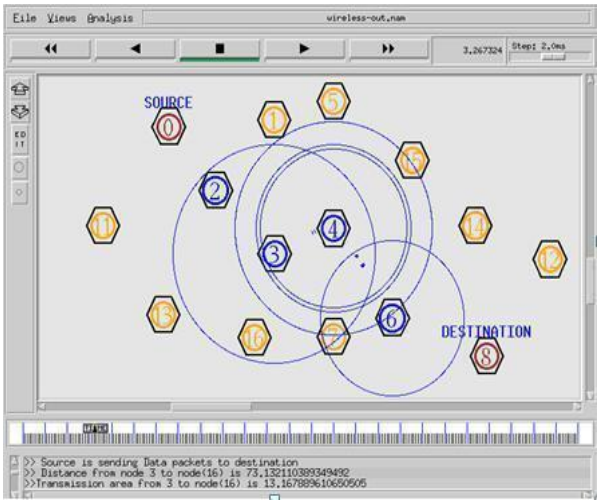


Fig 2: Network Topology

In the above network topology the nodes are moving randomly. After selecting the node 0 as source and node 8 as destination the next step is finding nearby relay node. After it checking the near relay node it again checks for shortest path from source to its destination.

This process continued up to find its destination path. After checking it transfer the packet from the source to its destination via throughput multi-hop. The hop is nothing but the gap between one node to another node.

In between the process of transferring the data packets to destination the nodes are moving randomly. According node movement the packets are transferred. In between the node is out of range by movement.

The Fig 3 shows that the node 4 is going out of particular range. So it again checks for nearby relay node. In the fig after node gone out of range, it searched and it takes the node 7 to transfer packets. So it doesn't have any connection with nodes to transfer any packets through that node 4.

So the remaining node doesn't have any connection with that type of nodes which are out of range.

If the node is out of range it again checks for the other nearby relay node. Up to there before nodes transfer packets and stores in queue. And if the nearby relay node came it again transfer those nodes.

So, by moving of nodes the out of range node may again come into the network. Then again it can transfer the before stored packets. And continued as nearby relay node as shown in Fig 4.

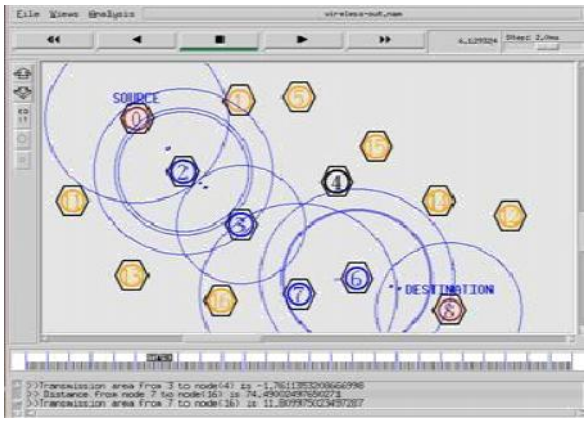


Fig 3: Network Topology

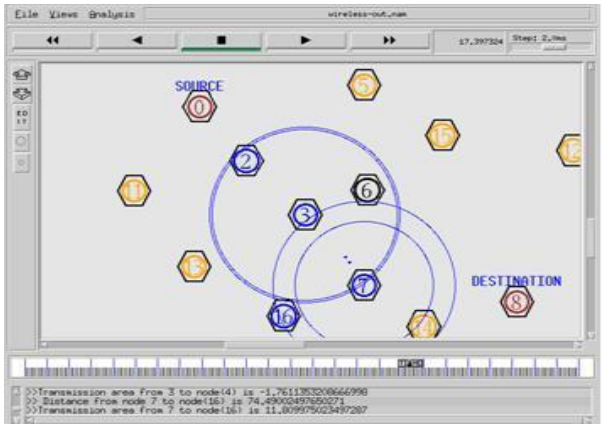


Fig 4: Network Topology

So it again checks the nearby relay node which is shortest path to its destination. And after that process it again sends the packets. If the previous node again comes to nearer to before nodes it again takes the previous send packets to its destination.

Multi-hop Relay Algorithm

In the above network number of nodes are 17. Multi-hop where having number of nodes which transfer a data packets to destination. Every node should have a queue to the packets.

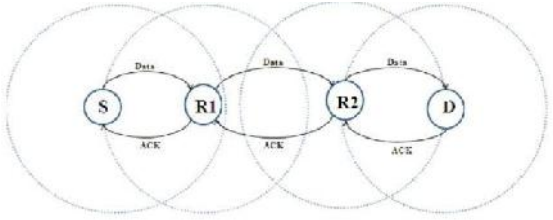


Fig 5: the Multi-Hop Relay.

Algorithm 1: Multi-hop Relay Algorithm

1. S checks whether its destination D is in the one-hop neighbourhood;
2. if D is within the one-hop neighbourhood of S **then**
3. S executes Procedure 1;
4. **else**
5. S randomly selects source-to-relay transmission or relay-to-destination transmission;
6. **if** S selects source-to-relay transmission then
7. S executes Procedure 2;
8. **else**
9. S executes Procedure 3;
10. **end if**
11. **end if**

Procedure 1: Source to Destination Transmission

1. S directly sends packet to D within 10% of time slot;
2. S waits for ACK within 90% of time slot;
3. **if** ACK not received **then**
4. S resends the packet to D;
5. **end if**
6. S deletes packet from its source queue;
7. D updates the Broadcast queue;

Procedure 2: Source to Relay Transmission

1. S randomly search for nearby relay node R out of one hop neighbours;
2. S directly sends packet to R within 10% of time slot;
3. S waits for ACK within 90% of time slot;
4. **if** ACK not received **then**
5. S resends the packet to R;
6. **end if**
7. S deletes packet from its source queue;
8. R updates the Relay queue;

Procedure 3: Relay to Destination Transmission

1. R search for destination node D;
2. R directly sends packet to D within 10% of time slot;
3. R waits for ACK within 90% of time slot;
4. **if** ACK not received **then**
5. S resends the packet to D;
6. **end if**
7. R deletes packet from its source queue;

Using the multi-hop relay algorithm the calculations of throughput and end-to-end delay, packet delivery ratio.

V. Simulation Results

Here based on location scenario under multi-hop the result are obtained in terms of throughput, end-to end delay and PDR.

End-to-End Delay Validation

The average time taken to transfer a packet to its destination. Delay of network = packet arrival time – sent time of connections. According to this formula the graph is plotted.

The number of nodes which are varied according to that network delay graph is plotted for multi-hop technique.

Packet Delivery Ratio (PDR)

Fig 6: The packet delivery to the destination

The above graph shows the packets delivered to particular destination within in the particular timeslot

When compared to previous technique the packet delivery ratio is high in multi-hop. By using this approach can delivery ratio can increase.

Packet Loss

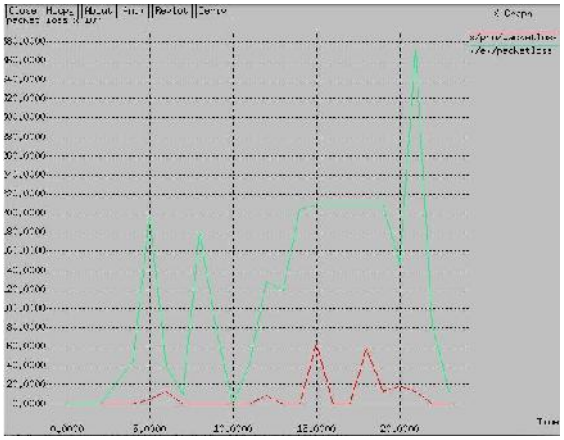


Fig 7: The packet loss between nodes during transferring to other nodes

The above graph shows when compared to previous the present technique packet loss is reduced. So, by using this technique can reduce the packet loss in between the nodes.

Routing overhead

The below graph shows when compared to previous technique the present technique routing overhead is less. So we can use the technique.

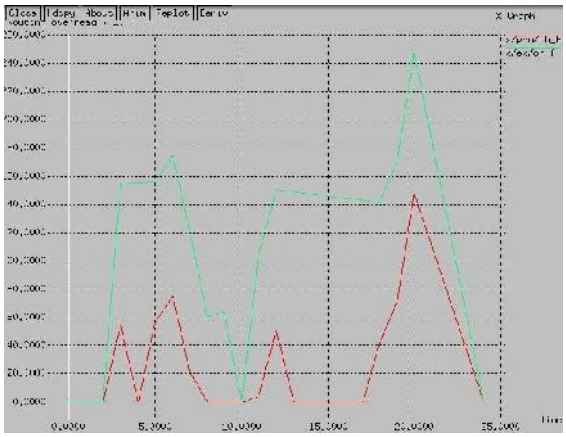


Fig 8: The Routing overhead graph

VI. CONCLUSION

Under The location based scenario using multi-hop relay algorithm when compared to previous relay algorithms the results shows better. Optimized delay by 2.23%, PDR 1.15% and average energy of network 10%. By moving of node if any link failure it again checking the nearby relay node and it transfer the packets to destination.

REFERENCES

[1] JingjingLuo, Jinbei Zhang, Li Yu and Xinbing Wang, " Impact of Location Popularity on Throughput and Delay in MANETs", May 2015.

[2] J.Burgess, B.Gallagher, D. Jehnsen and B. N. Levine, "MaxProp: Routing for Vehicle-based disruption-tolerant networking," Apr. 2006.

[3] P. Hui A. Chaintreau, J. Scott,R. Grass, J. Crowcroft and C. Diot, "Pocket switched networks and the consequences of human mobility in conference environments", Aug 2005.

[4] P. Juang, H. Oki, Y. Wang, M. Martonosi, L-S pehandD.Ruben-stein, "Energy-Efficient computing for wildlife tracking: Design Trade-offs and early experiences with ZebraNet" Oct 2002.

[5] A. Seth, D. Kroeker, M. Zaharia, S. Guo, S. Keshay, "Low-cost communiation for rural internet kiosks using mechanical backaul" Sept 2006.

[6] M. Grossglauser, D. N. C Tse, "Mobility increase the capacity of ad hoc wireless networks", Aug 2002.

[7] Michael J. Neely and Eytan Modiano "Capacity and Delay Tradeoffs for Ad-Hoc Mobile Networks" IEEE JUNE 2005.

[8] Grossglauser, Matthias "Mobility increases the capacity of ad-hoc wireless networks" IEEE, Nov 2007.

[9] Jiajia Liu, Xiaohong Jiang and Hiroki Nishiyama and Nei Kato, "Delay and capacity in adhoc Mobile networks with f-cast Relay Algorithms" IEEE, AUG 2011.

[10] Jiajia Liu, Xiaohong Jiang, Nishiyama, "On the delivery probability of Two-Hop Relay MANETs with Erasure coding", IEEE 2013.

[11] Amir Krifa, Chadi Barakat, *SeniorMember, IEEE* and Thrasyvoulos Spyropoulos, *Member, IEEE.*

[12] D. B. Johnson and D. A. Mallz, "Dynamic source routing in ad hoc wireless networks", in mobile comput, 1996.

[13] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing", 1999.

[14] Juntao Gaot, Xiahong, Jiangt, "Delay Modelling for Broadcasting-Based Two-Hop Relay MANETs", May 2013.

[15] J. Liu, J.Gao, X. Jiang, H. Nishiyama, N. Kato, "Capacity and delay of Probing-based two-hop relay in manets", IEEE, Nov 2012.

[16] Thomas kunz, Suranjit Paul, Li Li, "Broadcasting in Multihop wireless Networks: the Case for Multi Source Network Coding", IEEE, 2012.

[17] B. Williams, T. Camp, "Comparision of Broadcast techniques for mobile Ad hoc Networks", June 2002.

[18] Dipali K. Dakhole, Archana R. Raut, "Analysis of Multi-hop relay algorithm for efficient Broadcasting in MANETs".

[19] T. Kunz, S. Paul, L. Li, "Efficient Broad Casting in tactical Networks: Forwarding vs Network coding", Nov 2010.

Comparative Analysis of Shadow Detection and Removal Methods on an Image

V. Rashmi [1], V. Srinivasa Rao [2], K. Srinivas [3]

[1]*PG Scholar, Dept. Of CSE, V. R. Siddhartha Engineering College,* E-mail: rasvem12@gmail.com
[2]*Professor & Head, Dept. Of CSE, V. R. Siddhartha Engineering College,* E-mail: drvsrao9@gmail.com
[3]*Professor, Dept. Of CSE, V. R. Siddhartha Engineering College,* E-mail: vrdrks@gmail.com

***Abstract*— The shadow detection and removal is an important step in computer vision applications which has been a key challenge in various real life scenarios which are including under surveillance system, indoor outdoor scenes and tracking. Shadow detection and removal method should be implemented in indoor and outdoor with any objects like human, vehicles, and motorcycles moving objects in different times with different environments including weather, different sources of light and lighting conditions. Shadow detection after its removal is considered as the first step to shadow analysis and image processing in the number of applications. In this framework, recent techniques of shadow detection like Intensity based, Segmentation based, Mask Construction, Color based, Edge based methods are studied. Out of the shadow removal methods like Chromacity, Physical, Geometry, Small region texture based, Large Region Texture Based Method, the Otsu's thresholding along with Chromacity and the Geometry method have been discussed with their comparative analysis. Out of those studies the Otsu's Thresholding method is the best method for removal when compared to the other methods.**

***Keywords*-shadow, shadow detection, shadow removal, Otsu's thresholding, Chromacity based method, Geometry based method.**

I. INTRODUCTION

The influence of sunlight a shadow will be formed probably formed with respect to the moving object and which will result in wrong detection of the object, so that the result of a shadow is usually is mistaken for the object which effects of the moving object tracking. Thus it will be resulted to the incorrect identifying or analyzing the moving object. A significant role is played in refining of the vision of computer vision tasks including video surveillance, traffic monitoring and segmentation and tracking.

A. Shadow

The Shadow is an area where a direct light from the light source which cannot reach due to obstruction by an object due to the illuminated source. The shadow removal after its detection is an important in dealing with outdoor and indoor images. Removal of the moving object shadow should be in the direction of the moving object sequence considered as an important step in image processing.

A shadow is formed and appears on a region where light from a source cannot reach due to blockage which has been created by an object. Mainly, a shadows can be divided into two categories as shown in the figure 1.

The Shadows is of two types in its classification. The formation of the shadows are shown in the below figure 2 along with its types. They are self shadow or form shadow and cast shadow. The self-shadow is formed due to the part of an object which is not lit by a direct source of light. A cast shadow can be defined as the dim area which has been projected by the object on a surface. Cast shadow can be sub-divided into two types. They are called as umbra and the penumbra regions.

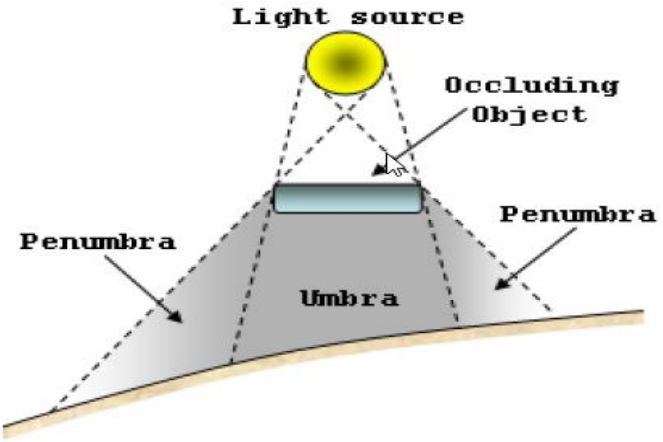


Fig. 1. Umbra and Penumbra formation of shadow [1]

The area where the direct light has been completely blocked is called the umbra region, and the part of the cast shadow in which the direct light is partially blocked is called as the penumbra region which will be formed due to the light source. The antumbra formation of the shadow is shown in the below fig. 2.

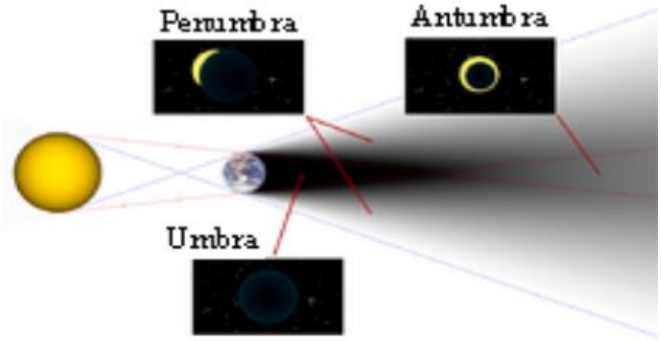


Fig. 3. Formation of shadows [2]

Hard shadows cause a loss of texture of the surface to a great extent of the shadow as shown in the below fig. 3.

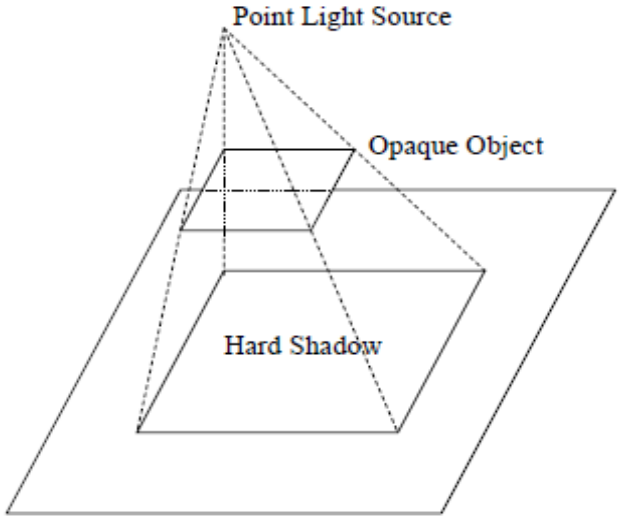


Fig. 3. Hard Shadow Formation

The soft shadows hold the texture of the surface on an image as shown in the below fig. 4 along with the penumbra and umbra formation with the light source through the opaque object.

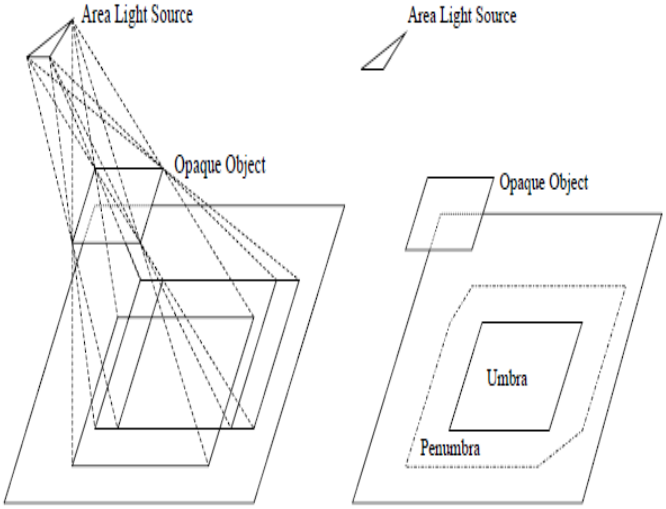


Fig. 4. Soft Shadow Formation

II. Different Types of Methods

To remove the shadow from the image there are different types of shadow detection approaches [3] and removal methods.

A. Shadow Detection Approaches

To remove the shadow [10] of the moving object firstly the shadow of the object must be identified. They are different shadow detection approaches [5] to detect the shadow [8]. Each method has its own advantage and disadvantages [4]. They are (i) Intensity, (ii) Texture, (iii) Segmentation, (iv) Mask Construction, (v) Color Based and (vi) Edge Based shadow detection approaches.

In the Intensity based shadow detection approach the cast shadow regions [8] will become darker as they are blocked by an illumination source. Depending upon the illumination the shadow can be much darker which initially reject the non-shadow regions to find out the range of the shadow regions. The standard deviation is calculated for the shadow pixel in the shadow detection approach based on intensity. The advantage here is the intensity information is directly calculated for the data and the disadvantage is the pixel intensity value is easily affected to illumination changes [7] of the moving object in the image.

The texture based shadow detection approach typically divided into two types of texture based shadow detection approaches. 1) The selection of the candidate shadow pixels or regions and 2) The classification of the candidate pixels or the regions either as foreground or shadow based on the texture correlation. The region under the shadow will be retained most of their texture.

Fig. 5. Shadow Removal form Texture Surface [2]

The fig. 5 shows the shadow removal from texture surfaces. The principle for the shadow detection based on texture information it differentiates the background, shadow and the foreground textures. This approach is best for indoor scenes and the texture capturing is difficult to implement and has poor performance for the outdoor scenes of an image.

In the segmentation based approach for the shadow detection principle is based mainly on the properties which has been possessed by the shadow pixels. The thresholding based method also comes under segmentation approach for shadow detection. It is simple and easy to implement because it can easily detect the probable shadow boundaries accurately but there is a chance of misclassifications of the shadows of the small objects by thresholding approach under segmentation approach.

In the mask construction for shadow detection uses the structuring element for the binary shadow mask to be computed. It gives the accurate results for the satellite images

in which this approach is computationally inexpensive and performance of outdoor scenes is better.

In the color based approach [3] color differences are used as the color tune value of a shadow and background of the object are of same but with different intensity. The color based approach is a reliable technique for color images. There is a main disadvantage here is it failed when an intensity of the shadow and the background are of same or if the color of the object is same or darker than the background.

The edge based [9] main principle is to detect the brightness changes sharply or the discontinuity and to detect the missing pixels. The edge detection approach gives the boundary between the shadow and the background of an image. The disadvantage here is the edge detection is not suitable for their small objects and their shadows.

The edge detection approach [9] gives the boundary between the shadow and the background.

A. Shadow Removal Methods

There are different shadow removal [6, 7] techniques [11,12] after the shadow detection [4]. They are Chromacity Based Method, Physical Method, Geometry Based Method, Small Region Texture Based Method, Large Region Texture Based Method.

The color spaces such as HSV, C1C2C3 and normalized RGB are the color models which have been proved to be robust for shadow detection in the Chromacity based method of shadow removal.

When an effect of the sky illumination increases, the shifting of the region of the shadow will be towards the blue component of an image in the Physical method of the shadow removal.

The orientation, shape and size of the shadow are predicted here in the geometry based method with the knowledge of the illumination source, object shape and the ground plane in the Geometry based method.

In the Small region texture based method, after selection of candidate shadow pixels, the classification of candidate shadow pixels as either foreground or shadow based on texture correlation of a shadow.

The Large region texture based method is not guaranteed significant textures so a method proposed using color features of the shadow is to first create large candidate shadow regions which have been discriminated from objects using gradient based texture correlation.

III. COMPARATIVE ANALYSIS

The comparative analysis of three different methods out of all the other methods such as Geometry Based Method and Chromacity Based Methods are done to analyze the performance with the Otsu's Thresholding Technique. Comparative evaluation is performed from the three test sequences. In the first evaluation step the four different sequences are compared with each other.

The original cast shadow image which has been given as input for shadow removal is shown in the fig. 5. This original image which has the cast shadow is taken as the input for the three required methods which includes the Otsu's Thresholding, Geometry based method and Chromacity based method [10].

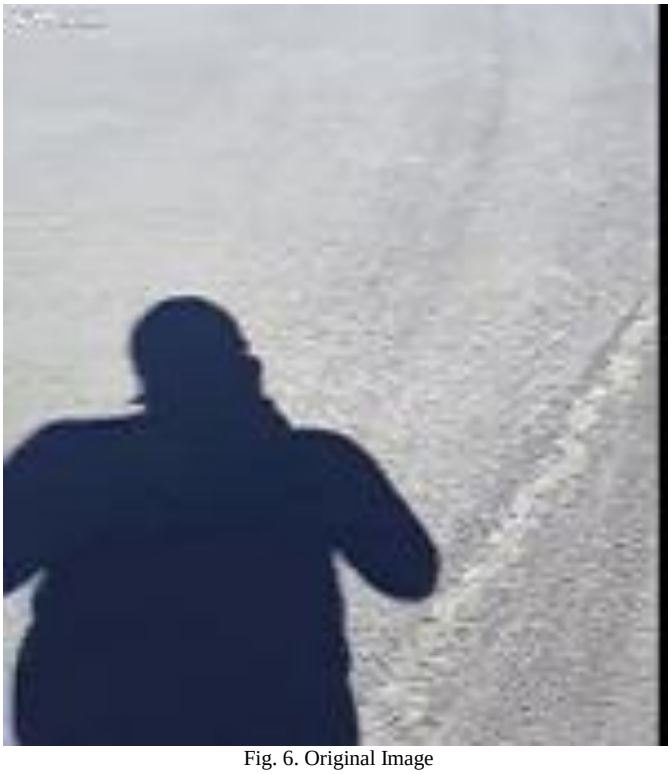


Fig. 6. Original Image

The Otsu's thresholding is considered for the cast shadow removal after detection of the shadow is shown in the fig. 7 as shown below from the fig. 6 which is the original image required to remove the shadow.

Fig. 7. Shadow Removed Image using Otsu's Thresholding

The Geometry based method of removal of the cast shadow for the original image is as shown in the fig 8.

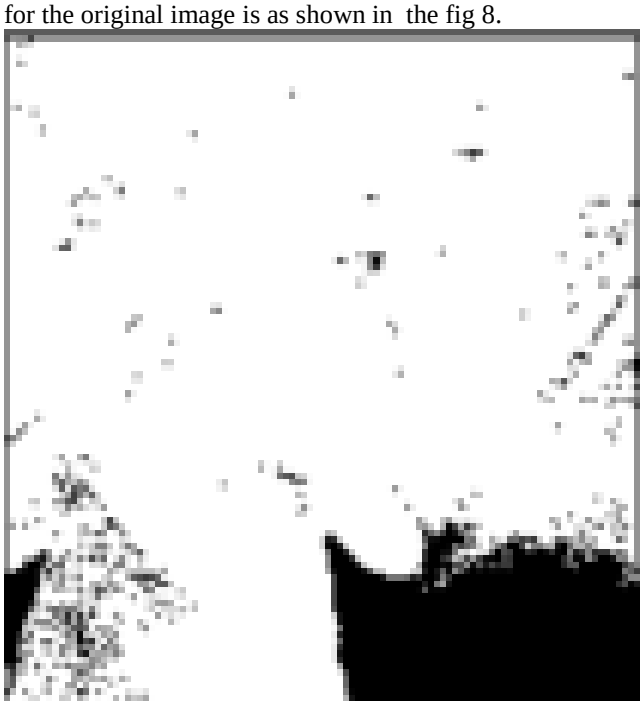


Fig. 8. Shadow Removed Image using Geometry Based Method

The Cromacity based method of removal of the cast shadow for the original image is as shown in the fig 9.

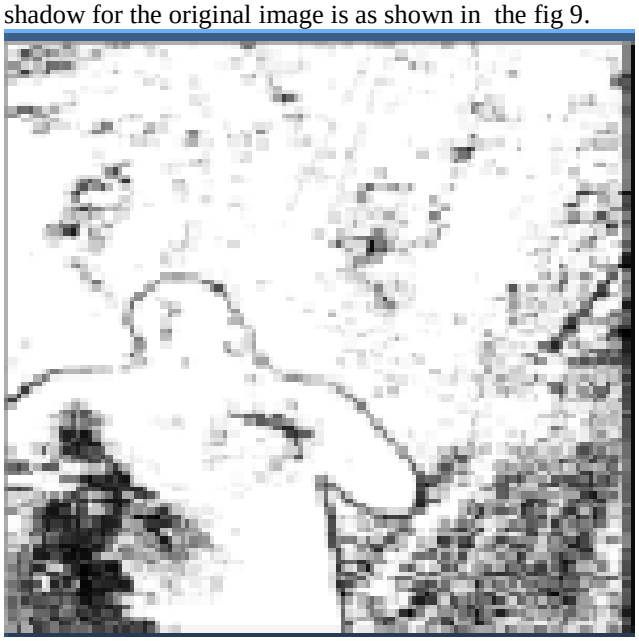


Fig. 9. Shadow Removed Image using Chromacity Based Method

Parameters required for the caliculation are the True Positives (tp), True Negatives (tn), False Negatives (fn) and False Positives (fp) which will determine the in recognizing the shadows. The False Positive is also known as the false allaram which evaluates the condition as true when the condition is false. False negative is where the test results declares the failure of the condition, mainly when it was successful. True positive means it has been correctly identified, such as the correct identification of the shadow and non-shadow regions. True negative is is equivalent to the correct recognition, in which it occurs when the predicted value and the acutual value are negative. Here, the classifier truely indicates that the non-shadow area is recognised as the non-shadow regions as shown in the table 1.

The sensitivity measures the performance of the binary classification test, which has been also known as the recall rate. It measures the function of the shadow as the actual positives which are been recognised as such.

Method Name	Shadow Removal			
	True Positive (tp)	False Positive (fp)	True Negative (tn)	False Negative (fn)
Otsu's Thresholding	8	1	9	2
Geometry	4	3	7	6
Chromacity	5	4	6	5

Table 1. Error Matrix

The specificity measures as the share of negatives which have been correctly recognised. The accuracy of the algorithm means the degree of the closeness of the measurements predicted by the algorithm to the actual value of the shadow.

The formula for the performance measures are shown below.

$$\text{Precision} = \frac{tp}{tp+fp} * 100\%$$

$$\text{Recall} = \frac{tp}{tp+fn} * 100\%$$

$$\text{accuracy} = \frac{tp+tn}{tp+tn+fp+fn} * 100\%$$

Error Rate = 1-accuracy

The table 2 shows the calculated results.

Category	Methods		
	Otsu's thresholding	Geometry	Chromacity
Total instances	10	10	10
Correctly removed	8	4	5
Incorrectly removed	2	6	5
Accuracy	0.85	0.55	0.55
Error Rate	0.15	0.45	0.45
Precision	0.88	0.57	0.55
Recall	0.8	0.33	0.5

Table 2. Performance Measures

Thus, from the above calculated performance measures show that the method used by Otsu's thresholding has better performance than the other two methods.

IV. CONCLUSION

In this paper the comparative analysis of various shadow detection and removal techniques has been presented along with the fundamentals of shadow, types of shadows and the

shadow detection and removal methods are observed. Each shadow detection and removal technique has its own advantages and disadvantages. Chromacity based method, Geometry based method and a method using Otsu's thresholding are compared with their respective outputs and which have been discussed with their comparative analysis. In this analysis, the Color based method and Otsu's thresholding are the efficient methods for shadow detection after its removal of an image. In all the methods, dark color region are considered as shadows like colored clothes are removed from the image. In the future work to overcome the problem.

REFERENCES

[1] Eli Arbel, "A Novel Approach for Shadow Removal Based on Intensity Surface Aproximation" March 2009.
[2] Jyoti Bala, Ritika, "Techniques and Algorithms of Shadow Detection in Images" International Journal of Core Engineering and Management, Volume 1, Issue 7, October 2014.
[3] Ariel Amato, Ivan Huerta, Mikhail G. Mozerov, F. Xavier Roca and Jordi Gonz`alez, "Moving Cast Shadows Detection Methods for Video Surveillance Applications" Volume 6, September 2012.
[4] M Jasmin T Jose, V. K. Govindan, "A Survey and Comparative Evaluation of Recent Methods" Internatioanl Journal of Computer Applications, Volume 45, No. 4, April 2013.
[5] Angie W. K. So, Kwan-Yee K. Wong, Ronald H. Y. Chung, and Francis Y. L. Chin, "Shadow Detection for Vehicles by Locating the Object Shadow Boundary" March 2009.
[6] Kaushik Deb, Animesh Kar, Ashraful Huq Suny, "Cast Shadow Detection and Removal of Moving Objects Based on HSV Color Space" on Smart Computing Review, Volume 5, No. 1, Feb. 2015.
[7] Jinhai Xiang, Heng Fan, Honghong Liao, Jun Xu, Weiping Sun, Shengsheng Yu, "Moving Object Detection and Shadow Removing under Changing Illumination Condition" Research Article, February. 2014.
[8] Salman H. Khan, Mohammed Bennamoun, Ferdous Sohel, Roberto Togneri, "Automatic Shadow Detection and Removal from a Single Image" IEEE Transactions on Pattern Analysis and Machine Intellengence, Volume 38, Number 3, March 2016.
[9] Maryam Golchin, Fatimah Khalid, Lili Nurliana Abdullah, Seyed Hashem Davarpanah, "Shadow Detection Using Color and Edge Information" Journal of Computer Science, March 2013.
[10] Priya Garg, Kirtika Goyal, "Detection and Removal of Shadow using Chromaticity", International Journal of Computer Science and Information Techniologies, Volume 5, March 2014.
[11] Savita Mogare, "A Survey on Various Shadow Detection and Removal Methods/Algorithms", International Journal of Recent Trends in Engineering and Research, Volume 2, Issue 3, March 2016.
[12] Divya S Kumar, Neenu Wilson, "A Review on Different Shadow Detection and Removal Methods", International Journal of Scientific Engineering and Applied Science, Volume 2, Issue 1, January 2016.

Comparative Analysis of Machine Learning Techniques on Stock Market Prediction

M. Sreemalli [1], P.Chaitanya[2], K. Srinivas [3]

[1]*Dept. of CSE, V.R Siddhartha Engineering College,* [1]E-mail:sreevalli.2340@gmil.com

[2]*Dept. of CSE, V.R Siddhartha Engineering College,* [2]E-mail:chaitu9876@gmail.com

[3]*Professor, Dept. of CSE, V.R Siddhartha Engineering College,* [3]E-mail:vrdrks@gmail.com

***Abstract*—Stock market prediction is a typical task to forecast the upcoming stock values. It is very difficult to forecast because of unbalanced nature of stock. Stock market prices are changing continuously. Many Stock holders invest more money on this stock market. Without having a clear idea on stock market, many people are losing a lot of money. More analyzing stock market may results in high profits. In this paper explore the use of artificial neural network is a very popular technique to forecast the stock market price and support vector machines. Other traditional methods are Hybrid markov model, Support vector machine and ARIMA models. Using these models to list the advantages and disadvantages of all these models and compare the performance of stock market.**

Index Terms:-
Artificial neural network, Support Vector Machine, Hidden Markov Model, optimize, stock market, Estimate, knowledge, accurately, Prediction.

I. INTRODUCTION

Stock market is a public market for the trading of a company stock and derivatives at an agreed price. In the stock market there is a trading between two share holders.It acquire share holders together to buy and sell their shares, and it sets the prices based on supply and demand.Whenever the share holders are buying their shares,based on demand,the stock prices are automatically increasing [2].If the company obeys all the listing requirements, they may issue their shares in the public.If the company has more than one stock exchange then it must be listed.These type of companies comes under the dual listing. In India, some companies are listed in National Stock Exchange (NSE) and Bombay Stock Exchange (BSE). Securities and Exchange Board of India (SEBI) must protect the share holders interests and to promote the development.

There are two important indicators for predicting stock price.This predicting analysis uses the data of company's financial reports, and technical information, and assumed that researching the trend in stock market.Using the fundamental analysis and technical analysis[1] can be used to analyze stock market [9].

II. PREDICTION TECHNIQUES

Presented the recent techniques in the stock market and give the comparative analysis of all these techniques.

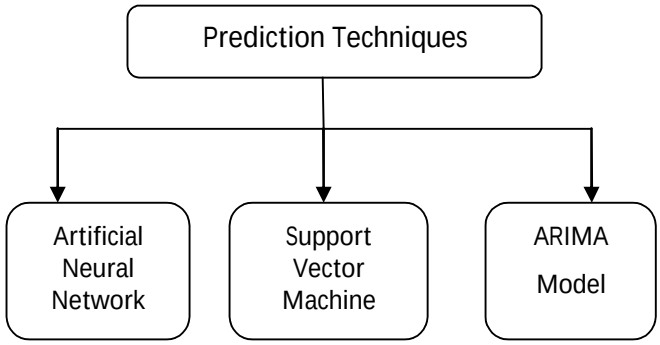


Figure 1: Prediction Techniques

A. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Network (ANN) is a mathematical model, it comes from the biological neural networks. Research is going on the ANN,it shows great potentiality on pattern recognition and machine learning problems such as classification and prediction. Artificial Neural Network is one of the machine learning approach which can handle discontinuous data, to predict the stock prices. Neural network is designed using certain number of neurons from the nervous system.

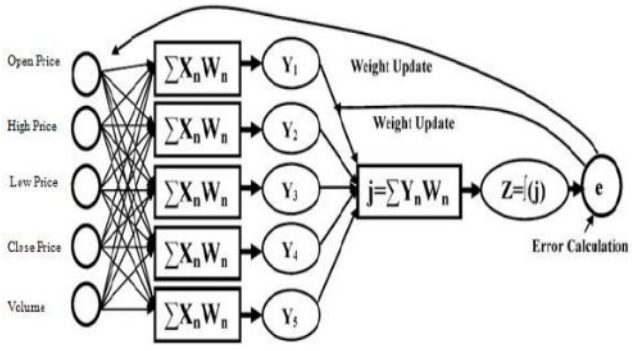


Figure 2. Artificial Neural Network Architecture[9]

At the hidden layer,it takes the input as output of the input layer and it performs the activation function,finally it produces the output,these output should be considered as a input of the output layer.In the output layer calculate the error.The actual and predicted value difference is called error rate.If the error value is not equal to zero,then again perform iterations upto the error value is zero [4].

ADVANTAGES:

- Artificial Neural Networks is one of the popular technique, it solves prediction problems .
- ANNs was used to solve many problems in financial time series forecasting.
- Using the neural network to predict the price with 90% accuracy.

DISADVANTAGES:

- Major problem in the Neural Networks is the Overtraining. The overtraining problem occurs by two main reasons, if use many nodes in the neural network, it takes more time to compute and long training time perio**d.**

C. SUPPORT VECTOR MACHINE:

The support vector classification (SVC) method used here, it has been proposed by Vapnik [5]. Using the linear model to implement the nonlinear class boundaries,which has been occur through some nonlinear mapping in which the input vector is fed into the high dimensional feature space. In the original space the nonlinear decision boundary is represented then the linear model is constructed in the new space. In the new space, an optimal separating hyper plane is constructed. So, the SVM is known as the algorithm that finds a special kind of linear model, it has the maximum margin hyper plane.

ADVANTAGE:

- Training the support vector machine involves optimization of a convex function with linear constraint. The problem has a unique global minimum which in turn overcome strucking to local minima observed in neural network, which reduces the computational cost.

DISADVANTAGES:

- SVM are more likely to avoid the problem of falling into local minimum.
- Support vector machine accuracy mostly ranges from59.35% and 71.43% only.

III. EVALUATION CRITERIA

To measure the performance of these techniques, computes the mean square error and root mean square error for these techniques, based on these techniques find out the error we find out good model. using the mean square error and root mean square error to compute the error value between actual value and target value.

$$MSE=\frac{1}{n}\sum_{i=1}(a_i - p_i)^2$$

$$RMSE=\frac{1}{n}\sqrt{\sum_{i\ 1}^{n}(a_i - p_i)^2}$$

In the above formula, where

a_i is the actual values

p_i is the predicted value.

C. ARIMA MODEL:

ARIMA model developed in 1970. It deals with timeseries data.ARIMA is called as dynamic and efficient model in time series forecasting which are especially used in shortterm predictions. The results obtained from this time series forecasting and explains the prospect strength of ARIMA models which are helpful to the investors for their decision making process. Gives to the Nifty bank dataset as input to the ARIMA model. The dataset consists of four elements are open, high,low and close prices.There is no significant pattern left in the auto correlationfunction.select the residuals from this function.These residuals of selected model are white noise.

ADVANTAGE:

- ARIMA models are resilient and efficient in time series forecasting ARIMA model has minimum standard error.

DISADVANTAGE:

- ARIMA model provides the short term prediction for stock investors, depends on that they take decision making process.

IV. EXPERIMENTAL RESULTS

NeuralNetwork:

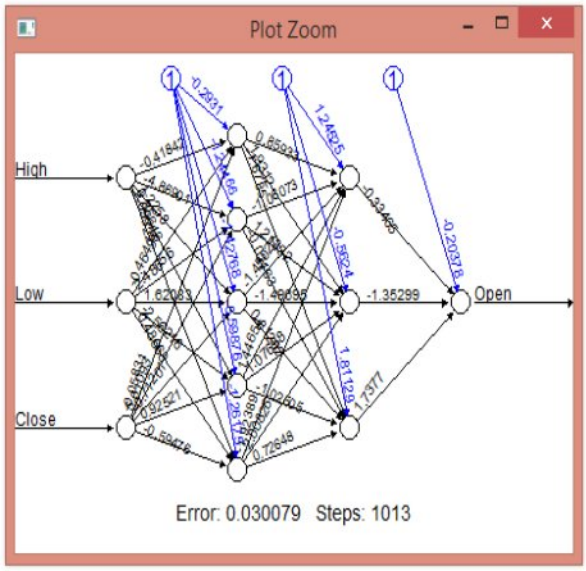


Figure 3: Neural Network Architecture

Here the neural network is created in R programming, Now the inputs consider here are High, Low and Close values.These inputs are fed in to the input layer and then one of the hidden layer have to be considered. The hidden layer containsfour nodes and one output layer.

Plot the Predicted Values:

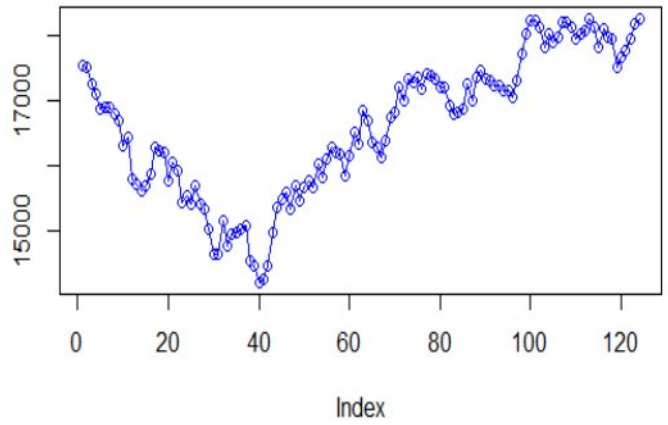


Figure 4: Predicted values

In the above figure, shows the predicted values for the Nifty bank dataset. The red color dots indicates the predicted values of dataset using the Neural Network.

Support Vector Machine:

Support vector machine is a technique which is used for estimating the relationships among the variables.It contains many techniques for analyzing the dependent and atleast one independent variable.This analysis helps to understand how these variables may vary ,that is if one variable is constant and the other can be change.

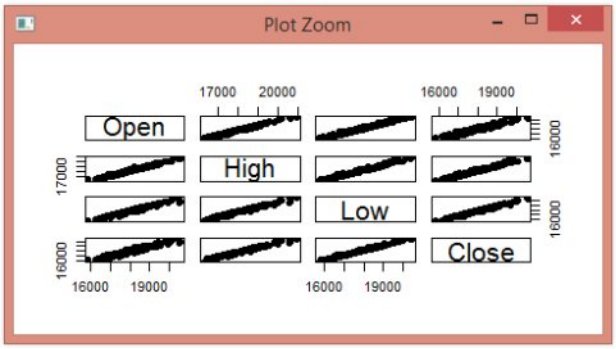


Figure 5:Plot the Open,high,low and close values

Using PCH option to plot the symbos when plotting points. These pch values are move from 0 to 25.For each number, there is a separate symbol to represent the points.

Plot the Predicted values:

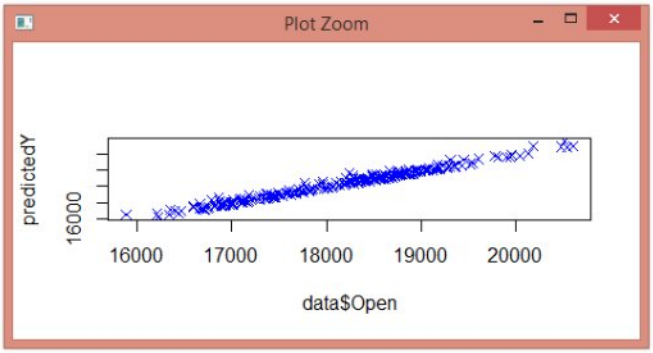


Figure 6:Plot the predicted values in graph

Plot the predicted values in the above figure. Predicts the open prices of future values and forecast the hundred days of data.these values are varies from 16000 to 20000.

ARIMA Model:

ARIMA are known to be robust and efficient models in financial and time series forecasting which are especially used in short-term predictions. The results obtained from this real-life data explains the potential strength of ARIMA models which are helpful to the investors for their decision making process.

Plot the Original Data:

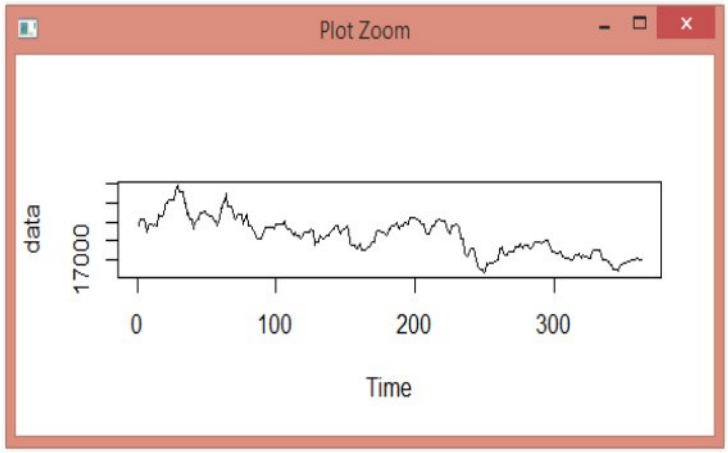


Figure 7: Plot the Original data

Plot the original data into the graph.In the above figure the x.axis line indicates the index values and the Y-axis line indicates the predicted values range.

Plot the Predicted values:

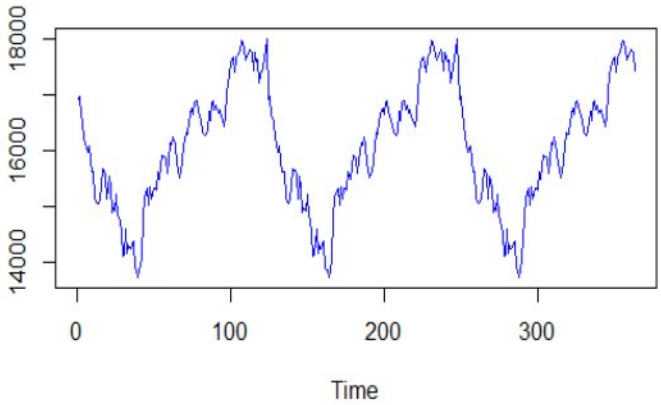


Figure 8: Plot the Predicted values

Forecast the predicted values using arima model and plot these predicted values in the above graph.

Error Curve:

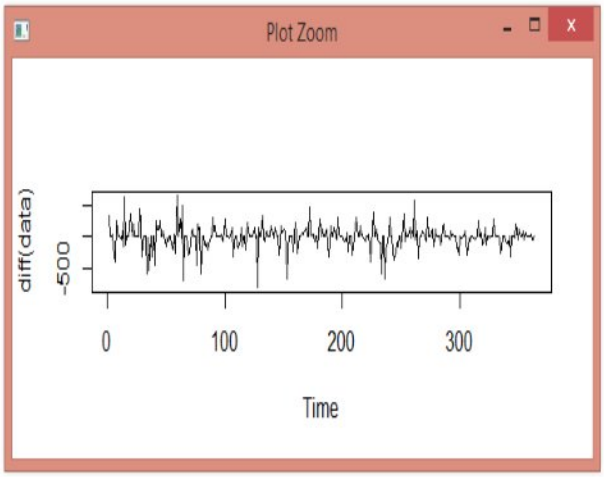


Figure 9: Error Curve

Plot the error values in the above figure. These error values are calculated from the actual and predicted values.

Performance Analysis:

Actual value	Predicted value (SVM)	Predicted value (NN)	Predicted value (ARIMA)	Error (SVM)	Error(NN)	Error (ARIMA)
16932	17119	17540	16824	187	608	108
16966	17156	17522	16575	190	556	391
16652	16842	17249	16474	190	597	178
16505	16693	17116	16377	188	611	128
16256	16440	16869	16049	65	613	207
16135	16321	16899	16104	186	764	31

Comparative Analysis of Machine Learning Techniques:

S. No	Techniques	Advantages	Disadvantages	Parameter used
1	Artificial Neural Network	Performance better than regression.It has lower predicion error	Prediction gets worse when noise variation is increased	Closing price of stock
2	Support Vector Machine	Does not lose much accuracy when applied to a sample from outside the training sample.	Can exaggerate minor fluctuations in the training data, thus resulting in decrease in subsequent predictive ability.	Net revenue, net income, price per earnings ratio of stock, consumer spending, diluted earnings per share, unemployment rate, consumer investment.
3	ARIMA	ARIMA models are resilient and efficient in time series forecasting ARIMA model has minimum standard error.	It is suitable for short term predictions only.	Open, high, low, Close prices and Moving Average.

V. CONCLUSION

In this paper, the use of machine learning techniques like, Artificial Neural network, Support Vector machine and Auto Regressive Integrated moving average for the prediction of Nifty bank data. Technical indicators are used to construct the relation between stock market index and their variables. Dataset used here is 2015 Nifty bank dataset. Implementing using Neural network consumes more amount of time in order to perform computations compared to other techniques, where as support vector machine has more error rate. Each technique has its own advantages and disadvantages. Different types of techniques have been used to predict the stock market and to forecast the future stock values up to some extent. Combining artificial neural network and Genetic algorithm may result in high accuracy.

REFERENCES

[1] Prashant S. Chavan,"Parameters for Stock Market Prediction",IJCTA,MAR 2013.
[2] Mayankkumar B Patel, Sunil R Yalamalle "Stock price prediction using Artificial Neural Network "IJIRSET, June 2014.
[3] Mrityunjay Sharma "Survey on Stock Market Prediction and Performance Analysis" IJARCET, January 2014.
[4] Gholap Rahul Mansing "Indian stock market prediction using neural network technique" IJARCET, March 2014.
[5] Shom Prasad Das, Sudarsan Padhy "Support Vector Machines for Prediction of Futures Prices in Indian Stock Market" IJCA, March 2014.
[6] Prakash Ramani Dr. P.D. Murarka "stock market prediction using artificial neural network "April 2013.
[7] G. Preethi, B. Santhi "Stock Market Forecasting Techniques: A Survey" JTAIT,December 2012
[8] Manasi Shah, Nandana Prabhu, Jyothi Rao "*P*erformance analysis of Neural *N*etwork Algorithms on Stock Market Forecasting" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 9 September, 2014.
[9] Gholap Rahul Mansing,"Indian stock market prediction using neural network technique" IJARCET, March 2014.
[10] Aditya Gupta and Bhuwan Dhingra, Non-Student members, IEEE, "Stock Market Prediction Using Hidden Markov Models," 2012.

A Relative Study on Open Source IaaS Cloud Computing Tools

Bala Savitha Jyosyula [#1], Suhasini Sodagudi[#]

[##]*Department of Information Technology, VRSEC, Vijayawada*
Andhra Pradesh,India.
[1]savitha1203@gmail.com

Abstract— **Service oriented architecture is a collection of services that communicate with each other to provide flexibility in system development and deployment. Cloud computing is an increasingly popular paradigm for accessing computing resources and providing services. There are various open source tools which provide these services. In this paper the comparison and study of various open source tools for IaaS such as Eucalyptus, Open Nebula, Nimbus and OpenStack have been discussed along with its architecture and implementation.**

Keywords: Open Source, Eucalyptus, Open Nebula, Nimbus, OpenStack

I. INTRODUCTION

IT specialist's constantly aims to have a rapid application development and deployment which is very difficult in this highly competitive and quick changeable universal business environment. There was a revolution in the computing of information society from distributed to cloud computing. Cloud Computing became a buzz word which is really something that appear to be a highly disruptive technology that is gaining a momentum. It is basically a form of distributed computing that allows users' ability to plug into a vast network of computing resources through the Internet. Cloud means **C**omputing **L**ocation independent **O**nline **U**tility that is available on-**D**emand. Major IT companies and academia give different definitions of the term "cloud computing" from different views but the definition given by the National Institute of Standards and Technology (NIST) is mostly comprehensive. NIST defines cloud computing as "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [1]".

Cloud computing is the integration of different services such as, Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS) and so on. IaaS provides the computing and storage capability on demand. PaaS provides a platform to receive Storage, networking and computing power on which the applications can be developed and executed. SaaS is using the developed applications. Different fundamentals of cloud computing are virtualization, scalability, interoperability, QOS, failure overcomes and cloud delivery models such as private, public, hybrid and community clouds. This paper mainly focuses on IaaS cloud open-source solutions.

A. Architecture of Cloud Computing

Cloud Computing architecture consists of front end and back end. The front end platform consists of fat client, thin client, mobile devices and back end platform consists of servers, storage, cloud based delivery and networks as shown in Figure 1.

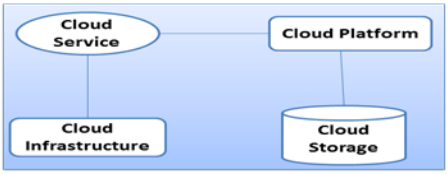


Fig. 1. Architecture of Cloud computing

The front end client platforms and cloud data storage communicates via middleware or web browser or using a virtual session. The cloud storage of the cloud architecture back end is deployed in the configurations of the public, private, hybrid and community clouds.

B. Document Overview

In section two, the background study is done about basics of cloud computing and IaaS. In section three the working of Infrastructure as a Service is known along with its characteristics. This section also gives a detail explanation of the architecture and implementation of the open source tools and summarizes the comparisons between those tools.

II. BACKGROUND STUDY

The following are some of the papers on cloud computing and establishing infrastructure as a service using open source tools. Cloud computing is the trending and emerging topic in these days .It is a model which provides access for shared pool of computing resources that can be rapidly provisioned.

Timothy Grance and Peter Mell in 2011 from NIST proposed the definition of the cloud computing and listed the five essential characteristics, four deployment models and three service models [1]. Eunmi Choi ,etal. in 2009 gave a detail about basics of cloud computing and different commercial and open source tools used to build services. [2]. Sushil Bhardwaj,etal. in 2010 provided a means of understanding and investigating IaaS. The authors also gave the outline of the the responsibilities of IaaS provider and the facilities to IaaS consumer [3]. Aniruddha S, etal. in 2013 discussed about the Infrastructure as a Service model and how to build it in cloud computing[4]. N. Nagar and U. Suman, in 2014 discussed about the architecture and

implementation of different open source tools and made a comparative analysis [5]. Amita and Rajender Nath in 2015 have explained the importance of open source tools in cloud computing and made a comparative study on some of the tools [6].

The background study helps in the survey of different open source tools for establishing Infrastructure as a Service with its architecture and implementations.

III. OPEN SOURCE IN CLOUD COMPUTING

As the usage of open source software has been increasing day to day, the open source software and cloud computing work together. Open source software runs at the bottom layer of the cloud and can also be used by different service models.

Open Source Cloud Computing [2] is a big blend of the cloud features such as lower service costs, economically affordable and having better resource sharing ability. The role this open source cloud computing is to build mechanism for identity management and to outline technological building blocks. Open-source cloud platforms make use of open-source hypervisors (KVM and Xen), but some of them also support commercial/closed hypervisors with exposed interfaces (VMware). Cloud platforms combine various tools of the underlying OS and virtualization layer with their own components in a more or less seamless cloud interface.

A. *Infrastructure as a Service*

Infrastructure as a Service is the delivery of the hardware such as server, storage and network with associated software such as operating system, virtualization technology and file system as a service. Without any long term commitment it allows users to provision resources on demand. The IaaS provider will generally provide the hardware and administrative services needed to store applications and a platform for running applications [3]. Characteristics and components of IaaS include:

- Utility computing services.
- Automating admin tasks.
- Scale in and scale out (dynamically).
- Virtualization.
- Policy-based services.
- Internet connectivity.

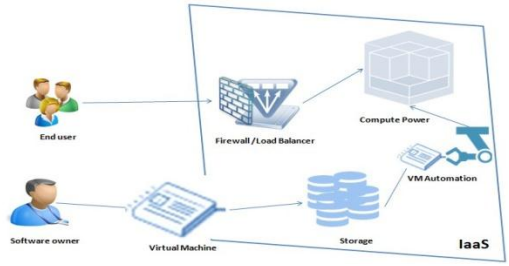


Fig. 2. Infrastructure as a Service

Running a user built virtualized machines can be provided by an IaaS. Figure 2 illustrates how a virtual machine is built, uploaded, configured, and then deployed for an IaaS environment. Using this technique virtual machines are created and loaded in the cloud with required software's. After the virtual machine is built it is uploaded to the hosting environment where it can be configured to use the raw storage. Once configured, the virtual machine can be deployed and started. Once the virtual machine is started it must ensure that the running virtual machine continues to work properly. IaaS provides a flexible option for migrating application to the cloud when there is no time to rework on the application's code [4].

B. *Open source Tools for IaaS*

Cloud tools can be used can provide the solutions for the organizational needs. Cloud tools implementation is considered as an important aspect [5]. The strategy of deployment may vary for each tool like the deployment can be performed through binary packages such as Cent OS, Open SUSE, Debian, and Fedora and deployed with UEC (Ubuntu Enterprise Cloud). Ubuntu 9.04 (Jaunty Jacklope) and any higher version of Ubuntu or UEC are required to deploy Eucalyptus. Ubuntu 10.04 or Cent OS 5.5 is highly recommended to install Open Nebula. Nimbus and Open Stack installed with any preferable Linux or Ubuntu version. Some of the open source tools for IaaS are discussed below:

1) *Eucalyptus*: Eucalyptus stands for Elastic Utility Architecture for Linking Your Program to Useful System. Eucalyptus was originated from the University of California at Santa Barbara, which is now supported by eucalyptus incorporation. It is free, open-source computer software for making Amazon Web Services (AWS) compatible for private and hybrid clouds. It enables pooling compute, storage, and network resources that can be dynamically scaled up or down as application workloads change. The Architecture of Eucalyptus mainly consists of five components as tabulated in [TABLE I.]

TABLE I
Components of Eucalyptus

COMPONENT	FUCTIONALITY
Cloud Controller (CLC)	Management of virtualized resources
Cluster Controller (CC)	Controls the execution of VMs
Walrus	Manages the storage system
Storage Controller (SC)	Provides block level network storage (supports Amazon-EBS)
Node Controller (NC)	Control VM activities (execution, inspection and termination of VM instances).

For the implementation of this tool a basic UEC (which manages euca2tools) need to be installed before deployment. Deployment spans into three servers where two servers run a 64-bit server version where, on server -1 CLC, SC, Walrus and CC will be installed and on server-2 hypervisors and node controllers will be installed and the third one runs a desktop 64-bit version as client. It can also be deployed using single server machine.

2) *Open Nebula:* The Open Nebula research project started in 2005 by Ignacio M. Llorente and Ruben S. Montero and the first public release of this software was on March 2008. It is a cloud computing platform which manages heterogeneous distributed data centers and a platform that manages virtual infrastructure to build private, public and hybrid IaaS clouds. The Architecture of Open Nebula mainly consists of five components as tabulated in [TABLE II.]

The Implementation of this tool can be done on the Ubuntu 10.04 or Cent OS 5.5 and the deployment can be performed through one of the Open Nebula's deployment tool Open Nebula Express. The installer of Open Nebula must be capable of deploying Ubuntu 10.04 – KVM – NFS, Ubuntu 10.04 – KVM – SSH, Cent OS 5.5 – Xen – NFS, Cent OS 5.5 – Xen – SSH and RHEL 5.5 – KVM – SSH. At least two nodes are required for the deployment of physical cluster nodes in this tool where one node work as a front end which runs all the Open Nebula services and the other nodes are treated as worker nodes.

TABLE II
Components of Open Nebula

COMPONENT	FUCTIONALITY
Interfaces and APIs	Management of physical and virtual resources
User and Groups	Access Control List for granting permissions.
Hosts and Virtualization	Runs on the server with installed hypervisors.
Networking	Provides network with support of VLANs and Open vSwitch
Storage and Image Repository	Repository of registered virtual machine images and supports both non shared and shared file systems.

3) *Nimbus:* Nimbus is the grouping of open source tools, which provides IaaS cloud computing solutions. By deploying resources on VMs it allows users to lease those resources remotely and by configuring them it helps to represent an environment required by a user. It is officially known as Virtual Workspace Service (VMS). It provides functionality to the users to use single nodes in clouds as a client or the user can launch auto configuring clusters or the user can build their own cloud using the workspace service. The architecture of the Nimbus consists of the components that are listed in the [TABLE III.]

For the implementation of this tool the user must first check the working of cloud configurations that serve remote users using the cloud-client, EC2 clients such as boto, and S3 clients such as s3cmd.The Configuration steps that involve in constructing VM are Service Dependencies *such as Sun Java 1.5 or later, python 2.5 or later (but not 3.x), Apache ant 1.6.2 or later and GCC*, Service Installation *such as service node, central services and image repository*, install DHCPd and configure networking.

TABLE III
Components of Nimbus

COMPONENT	FUCTIONALITY
Workspace Service	Allows clients to manage and administer VMs by providing two interfaces (Site Manager)
Workspace resource manager	Management of VM instance creation and implementation.
Workspace pilot	Makes necessary changes in site configuration and provides virtualization
Workspace control	Management of VM instance implementation such as start, stop and pause VM
Context broker	Provides Coordination for clients that allow launching large virtual cluster automatically and repeatedly.
Workspace client	Provides Complete access to Workspace service functionality.
Cloud client	Provides access to selected functionalities in the workspace service.
Storage service	Provides storage capabilities to store image.

4) OpenStack: OpenStack is the one of the top growing free open source software as well as a collection of open source software projects. It is an IaaS computing project jointly presented by NASA and the Rack Space. The Rack Space starting cloud formed in 2005 and it was rewritten in 2009. And again in 2010, they rewrite cloud servers and open source and released their first open source cloud computing tool Open Stack. NASA found the problem such as liability; scalability and so on in Eucalyptus and other tools. To overcome the underlying problems in tools, NASA decided to release its own object Nebula (not related with Open Nebula) in Feb 2010.The architecture of OpenStack includes different components as tabulated in [TABLE IV.]

TABLE IV
Components of OpenStack

COMPONENT	FUCTIONALITY
Horizon	Provides web interface (Dashboard) for administrators and users.
Nova	Management of VM instance and compute services
Keystone	Provides Identity service for authentication and authorization purpose.
Glance	Provides storage and retrieval of virtual machine images.
Swift	Provides Object Storage to run instances.
Neutron	Provides networking services that allow communication within virtual machine.

The implementation of the OpenStack can be mostly done on Linux based systems. The configuration of this component can be done in many ways but the basic way of the configuration is including component.conf (nova.conf) file, setting up the database, and integrating networking. Various steps involve in installing OpenStack are setting up an environment, setting up user, register an image, starting an instance and so on.

All the IaaS platforms have been designed to allow users to create and manage their own virtual infrastructures. However, these platforms have differences that need to be considered when choosing a platform. Some qualitative features [6] to consider as part of the selection are summarized in *[TABLE V.]*

TABLE V
Comparison of Open Source IaaS Tools

Feature / Property	Eucalyptus	Open Nebula	Nimbus	OpenStack
Computing Architecture	Hierarchical structure	Modular architecture	Three modules contain all the components	Message based architecture
Cloud Types	Private, Hybrid cloud	Public, private, Hybrid cloud	Public Cloud	Public, Private & Hybrid Cloud
Web Interface	CLI ,euca2tool and Web UI	Unix like CLI, Sunstone graphical interface	WSRF based or Amazon EC2 WSDL web interface	CLI ,euca2tool & NOVA API
Virtual machine manager	Xen, KVM , VMware(deprecated)	Xen, KVM and on-demand access to Amazon EC2	Xen, KVM, Bash, Libvirt	Xen, KVM
Live Migration	Not supported	Running VMs support	Traffic sensitive live migration.	Open virtualization format (OVF) support
Storage	Walrus (the front end for the storage subsystem), SAN for EBS	Nova, better Ceph support	Cumulus (Grid FTP and SCP)	Swift(Object storage), Cinder(Block storage)
Development Language	C, Java	C++, C, Ruby, Java, Shell script, lex, yacc	Java, Python	Python
Monitoring	With Nagios and Ganglia	Image, Template Repository Subsystem, Showback	Uses OpenTSDB	With Nagios, Zenoss
Load Balancing	Elastic load balancing(ELB) cloud controller	Nginx Server configured as load balancer	The context broker	Ironic Bare metal provisioning
Fault Tolerance	Separate clusters reduce the chance of correlated failures	Persistent database backend to store host and VM information	Checking worker nodes periodically and recovery	Use Swift
Uses	Geared toward persons interested in their cloud	Geared toward private company that want their own cloud.	Used in scientific Applications	Mission to produce ubiquitous cloud computing Platform
Scalability	Scalable	Dynamic Scalable	Scalable	Massively Scalable
Hypervisor support	VMware, KVM, Xen and ESX, Virtio	VMware, KVM, VirtualBox, Xen and libvirt	Xen 3.x or KVM and bash, ebtables, libvirt	Xen, KVM, Hyper-V QEMU, UML (User Mode Linux), XenServer, LXC
Reliability	Less Reliable	Less Reliable	Rollback host	More reliable as compare to others tools
Unique Features	Emulate Amazon AWS	Support for multiple types of users	Suited for small and midsized enterprises	Compatible and Connected

C. Summarization:

Organizations can set up their own cloud computing plan to fulfill business requirements. There are several clouds computing tools where most of them are based on open source and each tool have its own characteristics and advantages. There are various parameters that have been identified for the comparison i.e. Web Interface, virtual machine manager, live migration, Storage, Development Language, Compatibility, fault tolerance, load-balancing and monitoring. In this paper, Eucalyptus, Open Nebula, Nimbus and Open Stack are compared with respect to above mention parameters. Open Nebula and Open Stack support public, private and hybrid whereas Eucalyptus supports only private cloud. Nimbus supports public and private cloud. Eucalyptus provides storage compatibility with Elastic CC S3.The mention tools support KVM. Storage management provides storage capability of Eucalyptus through walrus, Open Nebula and Open Stack through Amazon S3 and so on.

IV. CONCLUSION

Open source cloud platforms provide flexibility, on demand services and allow great amount of customization. This paper focuses on the architecture and implementation issues of Eucalyptus, Open Nebula, Open Stack and Nimbus. It is found that OpenStack is suitable for rapid deployment of new products and Nimbus is well suitable for scientific community. Eucalyptuses, Open Nebula are suitable for private companies that want their own cloud.

The analysis and summarization done in this paper would help the users to understand the characteristics and would allow users to choose better services according to their requirements and also make more unified decision on the open source cloud platform according to their compatibility, interfaces and deployment requirement. By understanding some of the main differences between them, one can decide where and when each solution may be appropriate for its use.

REFERENCES

[1] Timothy Grance and Peter Mell, "The NIST Definition of Cloud Computing", *NIST Special Publication* 800-145, Sept. 2011.
[2] Eunmi Choi ,etal. "A Taxonomy and Survey of cloud Computing Systems" 2009 Fifth International Joint Conference on INC, IMS and IDC, IEEE, 2009.
[3] Sushil Bhardwaj, etal. "Cloud Computing: A Study Of Infrastructure As A Service (Iaas)" International Journal of Engineering and Information Technology, Vol 2,2010.
[4] Aniruddha S. Rumale, D.N.Chaudhari,"Cloud Computing: Infrastructure as a Service",*International Journal of Inventive Engineering and Sciences* ISSN: 2319–9598, vol-1, issue-3, 2013.
[5] N. Nagar and U. Suman, "Architectural Comparison and Implementation of Cloud Tools and Technologies" International Journal of Future Computer and Communication, Vol. 3, No. 3, June 2014.
[6] Amita and Rajender Nath "A Comparative Study of Open Source IaaS Cloud Computing Platforms" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 5, May 2015.

Ear Biometrics System based on Gray Level (Spatial) Statistical Feature Extraction

P.Ramesh Kumar
Department of Computer Science & Engineering
V.R.Siddhartha Engineering College
Vijayawada- 520 007, INDIA
E-mail: send2rameshkumar@gmail.com

SS Dhenakaran
Department of Computer Science & Engineering
Alagappa University
Karaikudi - 630 003, INDIA
Email: ssdarvind@yahoo.com

***Abstract*— In the world of surveillance monitoring, the need of passive human identification is increasing, where the human object is identified and verified for their identity. Since, Ear biometrics is a better suitable system where human subject ear image is obtained from surveillance video frame and processed to authenticate. Since the ear is a human's hearing sensor which is always visible to the camera. The proposed article extracts statistical features mean, median, mode, range, Standard deviation, Min, Max, Skewness and Kurtosis of the gray levels from the human ear image to construct an effective ear biometrics system.**

Keywords- Gray Level, Spatial, Statistical, Feature Extraction, Ear Biometrics

I. INTRODUCTION

Statistical feature of an input digital image provides a comprehensive view of gray or intensity information. The statistical parameter is also used for basic image enhancement, edge detection, restoration etc., In addition to basic image processing technique, the parameter can be used to construct feature database to verify a human. Image statistics give the information of the image intensity distribution analysis, the relationship between pixels and interpretation of uniformity of the pixels. The basic statistical parameter of the image can be considered as a good feature to differentiate one ear image from another by feature comparison in the extracted local ear statistical feature database.

The statistical features are based on the gray level intensity distribution of the input ear image and this information about the gray level distribution can be analyzed for understanding unique feature present in each ear image in the verification process. The input image can be processed for its first and second order derivatives and the texture features can be extracted from Gray Level Co Occurrence matrix. Biometrics are the development by identifying the authorization of an individual human based on physical parts such as finger, palm, Iris etc.; In general the biometric system can be divided based on active and passive involvement of the subject into the system.

The problem with the active biometrics system is subject cooperation, not hygiene and lost feature. To overcome the issues in the active biometrics system, the passive biometrics system is emerging where subject involvement is not required, the surveillance officer, get the required biometric physical image and extract feature to authenticate the individual human.

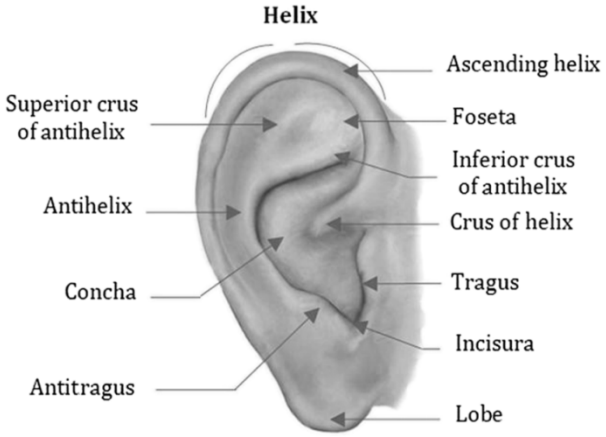


Figure 1: Overview of Outer Ear Anatomy [1]

The direction of research in biometrics is moving towards automatic recognition of the individuals, where ear biometrics can best adapt towards automatic recognition of the individuals in public places to stop unauthorized people. The ear prints local features are more stable and constant. The passive biometrics or automatic ear recognition, is the area where researchers working to achieve better recognition percentage. The next generation of ear biometrics is moving towards mobile based or automatic recognition based on surveillance camera. The primary difficulty associated with the ear biometrics system is the position variation of individuals with head and setting the ear image acquisition system, i.e., is the reason most of the ear biometric articles are based on a fixed set of ear database.

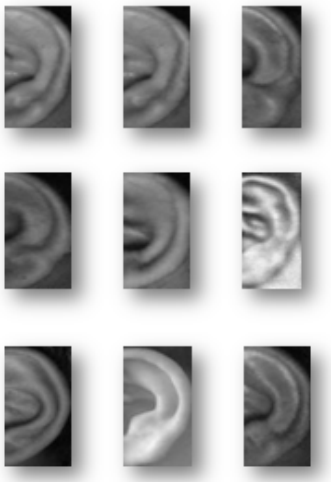


Figure 2: Ear Image from IIT Delhi Database [2]

II. METHODOLOGY

The statistical feature extraction methods for ear biometric system consists of the following phases

i. Enrollment
ii. Verification
iii. Identification

The general biometric system works in three phases: Enrollment, Identification and Verification. The biometric attribute information is gathered during the enrollment process and the collected features are stored in template storage (system disk). The subject present biometrics attribute is collected and evaluated with the feature template DB (created during enrollment) during identification (1: M or 1: N template comparison). During verification, the system, verifying that a subject is the individual that they claim to be, based over validating a biometric attribute gathered from the same individual (1:1 template comparison).

The enrollment is the process of registering the individual's ear image features into the local database for future verification. The enrollment process consists of the following stages

1. Image Acquisition (using ear DB)
2. Preprocessing (Quality Check)
3. Statistical Feature Extraction.
4. Creating the Feature Template DB storage.

The verification process authenticates the individual ear image feature with a single feature, database template (1:1). The verification process consists of the following steps

1. Capture the probe Ear Image
2. Preprocessing (Quality Check)
3. Statistical features extraction methodology
4. Verify the claimed identity using Matcher algorithm (1:1)
5. Claimed Identity is True/False.

The Identification process authenticates the individual ear image feature with a collection of local feature database (1: M). The Identification process consists of the following steps

1. Capture the probe Ear Image
2. Preprocessing (Quality Check)
3. Statistical features extraction methodology
4. Identify the User feature Template from M feature template DB using Matching Algorithm
5. User's identified or not identified.

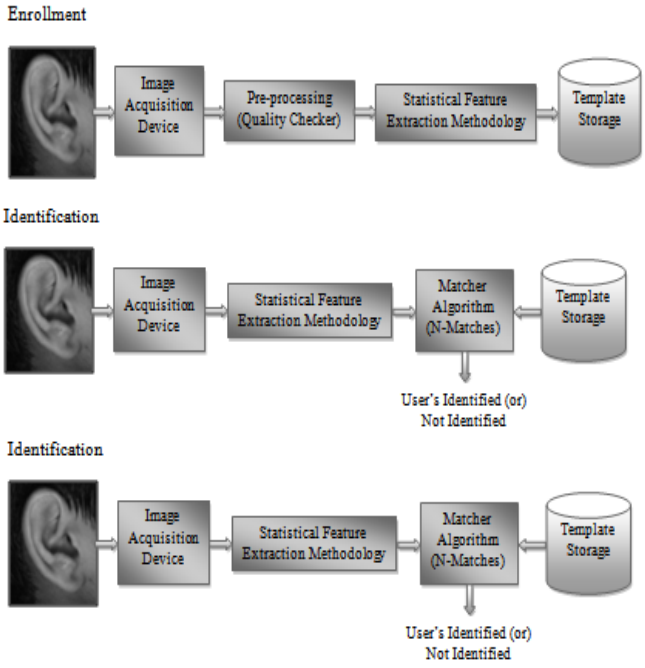


Figure 3: The process of Enrollment, Verification and Identification[3]

The sample collection can be downloaded from the existing ear database from IIT Delhi, India which consists of preprocessed, normalized and cropped ear images of size 50 x 180 pixels of 212 users with 754 ear images [2].

The statistical features are based on the gray level intensity distribution of the ear image. The working ear biometrics model can be built by extracting the ear image randomly from the moving objects in the surveillance video frame. The captured ear image of the subject can be verified by the proposed method as shown in the figure, but the proposed article we have used an existing database from IIT Delhi for our experimentation.

The initial input ear image is an RGB color image and it is converted into a gray level image and 11 different statistical features are extracted for verification. A gray level ear image represents a range of gray levels from 0 to 255 values with respect to the specific RGB range at a pixel point. The lowest

possible gray value is 0 (white) and highest possible value is 255 (black). The statistical features extracted in the proposed article are Mean_GL, Variance, Standard Deviation (SD), Max_GL, Min_GL, Range_GL, Median_GL, Mode_GL, Entropy (e), Skewness and Kurtosis.

The above work can be extended for texture feature of an ear image; Where GLCM (Gray level Co- occurrence Matrix) seems to be a good statistical approach for extracting texture feature, where texture defines the characteristics of an image. GLCM quantify the gray level distribution of pixels to its neighborhood pixel. The construction of GLCM depends on the relationship between two neighboring pixels. If i and j are the gray levels of pixel points p(x, y) and p(x+1,y+1) then co-occurrence matrix elements provide the difference in the gray levels of p(x, y) and p(x+1,y+1) with a distance 'd' on the image and certain texture feature such as Contrast, Correlation, Energy, Homogeneity, Kurtosis and Skewness can be extracted from a constructed GLCM Gray level Co-occurrence matrix.

III. STATISTICAL FEATURES

Mean_GL: The mean is an average gray level in the sample ear image I.

$$Mean\_GL = \sum_{i=0}^{M} \sum_{j=0}^{N} \frac{I(i,j)}{MN} \quad (1)$$

Where

Mean_GL - Mean of the Ear Image I
MN - No. of rows/Columns in the Image I
I (i, j) - Intensity Value at (i, j)

Variance (σ^2): The Variance of the ear image I provide overall intensity variance in I.

$$\sigma^2 = \mu_2(r) = \sum_{i=0}^{L-1} (r_i - m)^2 . P(r_i) \quad (2)$$

Where

σ^2-Intensity variance in Image I
μ_2- Second moment of 'r'about its mean
r - Discrete random variable representing Intensity level range [0,L-1]
m – Mean of Image I

Standard Deviation (SD): Both SD and Variance is the measure of contrast in an image I.

$$SD = \sqrt{\sigma^2} \quad (3)$$

Where σ^2-*Variance of Image I*

Min_GL (Min Gray level): Extracting the smallest gray level elements in the given image I.

$$Min_{GL} = \min_{(x,y)\in I(i,j)} \{f(x,y)\} \quad (4)$$

Max_GL (Max Gray Level): Finding the largest gray level element in the given image I.

$$Max_{GL} = \max_{(x,y)\in I(i,j)} \{f(x,y)\} \quad (5)$$

Range: Range provides the difference between largest gray level and smallest gray level in the sample ear image I.

$$Range(R) = Max_{GL} - Min_{GL} \quad (6)$$

Median_GL: The median gray level provides the center gray level which can be identified by arranging the gray levels in ascending order then get the middle value of the ear image I.

Mode_GL: The mode provides gray level which is repeated more than any other gray level in ear image I.

Entropy (e): Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image.

$$e = -\sum (P(r_i).log_2(P(r_i))) \quad (7)$$

Where

$P(r_i) = \frac{r_i}{MN}$ *Probability of occurrence of* r_i *i=0, 1, 2…….. (L-1)*

Skewness: Skewness is a measure of symmetry or more precisely the lack of symmetry or asymmetry.

$$Skewness = \frac{(1-p)}{SD} \quad (8)$$

Where

p = 0.4
SD - Standard Deviation

Kurtosis: Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution.

$$Kurtosis = \frac{(1-3p)+3p^2}{V} \quad (9)$$

Where

p=0.4
V – Intensity Variance

IV. PERFORMANCE

The general biometric performance measures are the False Acceptance Rate (FAR) and False Rejection Rate (FRR). The FAR and FRR metrics provide the probability of invalid user input parameter which are incorrectly accepted and probability of valid user inputs which are incorrectly rejected. The other biometrics performance metrics are CER: Crossover Error Rate, FER: Failure to Enroll Rate, Speed and Number of Template; [4][5]

Performance evolution defines the performance of the ear images features present in the database or not. The, Genuine accept rate is defined as, we check the true image features present in the database or not, if the result is true it is genuinely accept rate, and if the result is false is known as false reject rate.

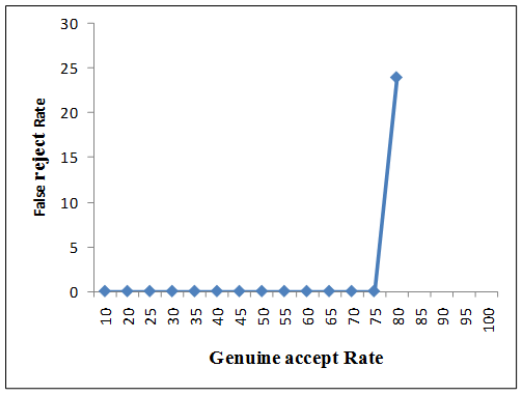


Figure 4: True images feature present in the database or not

The figure shows the performance evaluation of the given database. It shows that out of 100 images, 75 images are accepted by the database with genuine accept rate and remaining 25 are rejected with false reject rate. We take genuine accept rate in x-axis and false reject rate in the y-axis. Here, we check whether the false image feature present in the database or not. If the false images present in the database, it is false accept rate otherwise it is false reject rate.

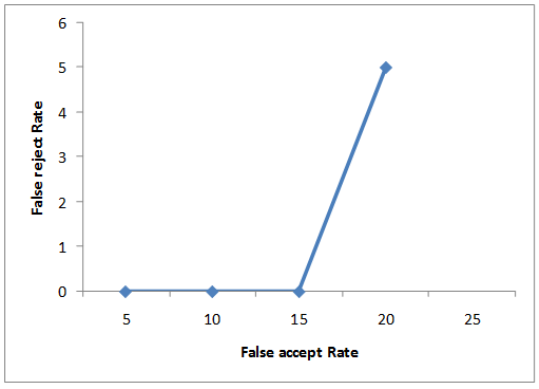


Figure 5: False images feature present in the database or not present

The above figure shoes that we take twenty images of the performance evaluation from the database. After evaluating the performance with the database, it shows that out of 20 images, 15 images are accepted by the database known as false accept rate and the remaining 5 images are not accepted. Those are rejected by the database called as false reject rate as shown in graph 4. We consider false accept rate in x-axis and false reject rate in the y-axis.

V. CONCLUSION

In this article we have proposed a statistical based feature extraction method to develop an ear image based human identification system. The experimentation was conducted based on IIT Delhi ear database which consist of 726 processed ear images. The proposed approach uses gray level distribution and its properties of the ear image to extract statistical features to build a biometrics system. The features mean, median, mode, range, Standard deviation, Min, Max, Skewness and Kurtosis extracted on the IIT Delhi ear database and the feature DB is created to verify the authorization of the individual. The performance is verified using False Acceptance Rate (FAR) and False Rejection Rate (FRR).

TABLE I. TABLE TYPE STYLES

S.NO	File Name	Mean	Variance	STD	Min	Max	Range	Median	Mode	Entropy	Skewness	Kurtosis
1.	001_1.bmp	108.7233	1073.286	32.761	3	186	183	116	119	6.7788	0.0183	0.0003
2.	001_2.bmp	106.913	962.3348	31.0215	4	183	179	115	129	6.6462	0.0193	0.0003
3.	001_3.bmp	105.9476	1079.381	32.8539	3	176	173	114	117	6.7321	0.0183	0.0003
4.	001_4.bmp	105.37	1055.768	32.4926	4	176	172	113	131	6.7287	0.0185	0.0003
5.	001_5.bmp	110.0152	1170.495	34.2125	2	184	182	117	123	6.8397	0.0175	0.0002
6.	001_6.bmp	109.6127	1135.538	33.6978	4	185	181	116	117	6.8309	0.0178	0.0002
7.	002_1.bmp	89.7147	720.3342	26.839	5	171	166	93	97	6.598	0.0224	0.0004
8.	002_2.bmp	93.1137	753.3356	27.447	4	182	178	96	93	6.6239	0.0219	0.0004
9.	002_3.bmp	94.9817	660.7743	25.7055	4	185	181	97	91	6.5631	0.0233	0.0004
10.	003_1.bmp	107.4474	924.4306	30.4045	3	178	175	114	124	6.6232	0.0197	0.0003
11.	003_2.bmp	108.6203	910.6716	30.1773	5	172	167	115	116	6.6461	0.0199	0.0003
12.	003_3.bmp	105.9658	1040.239	32.2527	4	179	175	113	120	6.7138	0.0186	0.0003
13.	003_4.bmp	105.6577	1010.751	31.7923	2	178	176	113	116	6.7046	0.0189	0.0003
14.	003_5.bmp	110.1207	846.3688	29.0924	4	174	170	117	122	6.6446	0.0206	0.0003
15.	004_1.bmp	73.6054	501.7466	22.3997	4	144	140	75	81	6.3632	0.0268	0.0006
16.	004_2.bmp	72.5038	544.9968	23.3452	0	147	147	74	61	6.4373	0.0257	0.0005
17.	004_3.bmp	76.6787	553.4536	23.5256	3	164	161	76	66	6.4606	0.0255	0.0005
18.	004_4.bmp	76.5494	621.267	24.9252	2	162	160	77	71	6.5079	0.0241	0.0005
19.	004_5.bmp	74.178	647.7258	25.4505	4	149	145	80	87	6.4726	0.0236	0.0004
20.	004_6.bmp	74.337	655.6981	25.6066	3	151	148	80	94	6.4876	0.0234	0.0004
21.	005_1.bmp	57.104	460.5121	21.4595	3	124	121	58	51	6.3316	0.028	0.0006
22.	005_2.bmp	62.5762	462.8483	21.5139	2	129	127	62.5	60	6.3961	0.0279	0.0006
23.	005_3.bmp	64.7432	498.1431	22.3191	3	132	129	65	65	6.4341	0.0269	0.0006
24.	005_4.bmp	62.3023	578.3939	24.0498	5	137	132	64	58	6.4958	0.0249	0.0005
25.	006_1.bmp	135.3161	861.4917	29.3512	61	191	130	138	107	6.6619	0.0204	0.0003

REFERENCES

[1] Ear biometric recognition using local texture descriptors Amir Benzaoui ; Abdenour Hadid ; Abdelhani BoukroucheJ. Electron. Imaging. 23(5), 053008 (Sep 19, 2014). doi:10.1117/1.JEI.23.5.053008

[2] Ajay Kumar and Chenye Wu, "Automated human identification using ear imaging," *Pattern Recognition*, vol. 41, no. 5, March 2012.

[3] Davide Maltoni Tutorial Presentation on " Fingerprint Recognition: Basics and Recent Advances " *Biometrics (ICB), 2012 5th IAPR International Conference on*. IEEE, 2012

[4] http://www.biometricsolutions.com/index.php?story=performance_biometrics

[5] https://en.wikipedia.org/wiki/Biometrics#Performance

[6] Devesh Narayan, Sipi Dubey 'A Survey Paper on Human Identification using Ear Biometrics', International Journal of Innovative Science and Modern Engineering (IJISME), ISSN: 2319-6386, Volume-2 Issue-10, September 2014.

[7] Anika Pflug, Christoph Busch, 'Ear Biometrics-A Survey of Detection, Feature Extraction and Recognition Methods', IET Biometrics, Volume 1,Number 2, pages 114-129, June 2012.

[8] Jain, Anil K., Arun Ross, '*An Introduction to Biometric Recognition'*.IEEE TRANSACTIONS on circuits and systems for video technology 14.1 (2004): 4-20. IEEE Explore. Web. 5 Dec. 2011.

[9] Abaza, A., Ross, A., Herbert, C., Harrison, M. A. F., and Nixon, M. S. 2011. 'A Survey on Ear Biometrics'. ACM Trans. Embedded. Comput. Syst. 9, 4, 33 pages, Article 39, March 2010.

[10] D. J. Hurley, B. Arbab-Zavar, and M. S. Nixon, 'The Ear as a Biometric', In A. Jain, P. Flynn, and A. Ross, Handbook of Biometrics, Chapter 7, Springer US, 131-150, 2007.

[11] Michal Choras, 'Image Feature Extraction Methods for Ear Biometrics'. University of Technology & Life Sciences, Bydgoszcz. Computer Information Systems and Industrial Management Applications, 2007. IEEE Explore. Web. 28-30 June 2007

[12] Hanna-Kaisa Lammi, '*Ear Biometrics'*, Department of Information Technology, Lappeenranta University of Technology, Laboratory Information Processing, Lappeenranta, Finland, 2004.

[13] A Survey on Human Ear Recognition System Based on 2D and 3D Ear Images', Durgesh Singh, Sanjay K. Singh, Department of Computer Science and Engineering, Indian Institute of Technology

[14] AMI ear database', Esther Gonzalez, Luis Alvarez and Luis Mazorra, PhD, Department of computer science and technology, under a Creative Commons Reconocimiento- No commercial- SinObraDerivada 3.0.

[15] Hurley, David J., Mark S. Nixon, and John N. Carter. "Automatic ear recognition by force field transformations." Visual Biometrics (Ref. No. 2000/018), IEE Colloquium on. IET, 2000.

[16] Yang, Mingqiang, Kidiyo Kpalma, and Joseph Ronsin. "A survey of shape feature extraction techniques." Pattern recognition (2008): 43-90.

[17] Sahoo, Soyuj Kumar, Tarun Choubisa, and SR Mahadeva Prasanna. "Multimodal biometric person authentication: A review." *IETE Technical Review* 29.1 (2012): 54-75.

[18] Jeng, Ren-He, and Wen-Shiung Chen. "Two Feature-Level Fusion Methods with Feature Scaling and Hashing for Multimodal Biometrics." *IETE Technical Review* (2016): 1-11.

Mr.P.Ramesh Kumar. B.Tech(CSE), M.Tech(CSE)., working as Sr.Assistant Professor in the Department of Computer Sciecne & Engineering ., V.R. Siddhartha Engineering College, Vijayawada, INDIA. His research interest include Biometrics, Image and Video Processing, IoT-Internet of Things. He has published 12 research papers till now in various International Conferences, Proceedings and Journals.

Dr. S.S. DHENAKARAN PhD., working as Professor in the Department of Computer Science and Engineering Alagappa University Karaikudi – 630003 Tamil Nadu, INDIA. His research interests include Cryptography, Mathematical Algorithms, Data Ming Image Processing, Data Mining and Network Security. He has published 40 research papers till now in various National, International Conferences, Proceedings and Journals.

Person Re-Identification across Multiple Camera Views

V Ramya[#1] , V V Vineela[#2],V Srinivasa Rao[#3] ,K Srinivas[#4]

[#]*Department of Computer Science and Engineering, VR Siddardha Engineering College, Vijayawada, Andhra Pradesh, India.*
[1] ramyarupav@gmail.com
[2]venkatavineela3@gmail.com
[3]drvsrao9@gmail.com
[4]vrdrks@gmail.com

[*]*Software Group, Advanced Data Processing Research Institute– ISRO, Hyderabad, Telangana, India.*

Abstract*---Person re-identification is an approach of identifying the person at different locations and time across camera views in surveillance system. Person re-identification probably the open challenge for low-level video surveillance in the presence of a camera network. As person move from one camera view to another camera view the same person countered as two different persons .In order to reduce false count and enable seamless tracking, we proposed model-free gait representation. In this approach, width vector profile and width vector mean are taken as features. Applying of Normalization on features helps to achieve the results more accurately. To solve classification problem different distance metrics are used. The Experiments are carried out on CASIA gait database of gait dataset A. Re-identification results provided for the normalization of features and without normalization of features, results recorded for different view angles with respect to camera and results of applied different distance metrics are provided.

Keywords—person re-identification, width vector profile, width vector mean, normalization.

I. INTRODUCTION

Person Identification or recognition has been receiving broad interests and it is highly desirable in applications such as security monitoring, authentication, etc. In order to recognize a person, different traits, including fingerprint, face, iris and gait can be used. Among these possible traits, face and body are preferred since they can be acquired without the person's cooperation.

Using of human operator manual re-identification in large camera networks are expensive and inaccurate. Now a days many people using Gait recognition system since it has unique advantages as compared with other biometrics. Gait recognition is a task to identify or verify Indiduals by the way they walk shown in Fig.1.In video surveillance based application identifying the human gait is Important because it captures the human from a distance. So we choose gait based approach for person re-identification.

Fig.1. Gait Cycle

Gait recognition methods mainly classified into two major methods [1]; model-based and model-free methods. Model-based methods obtain series of static or dynamic body parameters via modeling or tracking body components such as limbs, legs, arms and thighs. View-invariant and scale-independent are main advantages of model-based approach. But model-based approaches are sensitive to the quality of gait sequences to achieve high accuracy and their computational cost also high due to it's large parameter calculations. Model-free approaches focus on either shapes of silhouettes or the whole motion of human bodies. Model-free approaches are insensitive to the quality of silhouettes and also have the advantage of low computational costs.

A.Overview of the proposed method

The structure of proposed gait recognition system is in Fig.2.The system consist of mainly of three modules: One is preprocessing unit, here we detect the human movements and then extract the binary silhouette image from each frame. Background is eliminated from each image. The second feature extraction unit, extract the width vector features from normalized silhouette images. In the third step for the

classification different distance metrics are applied for person re identification. This paper organized as follows. In Section 2, describes the related work which was done up to now. Section 3, describes the proposed methodology of our system.

Section 4, describes the algorithm that we performed. Section 5, describes the Experimental results presented on CASIA database. Section 6, describes the conclusion of this paper.

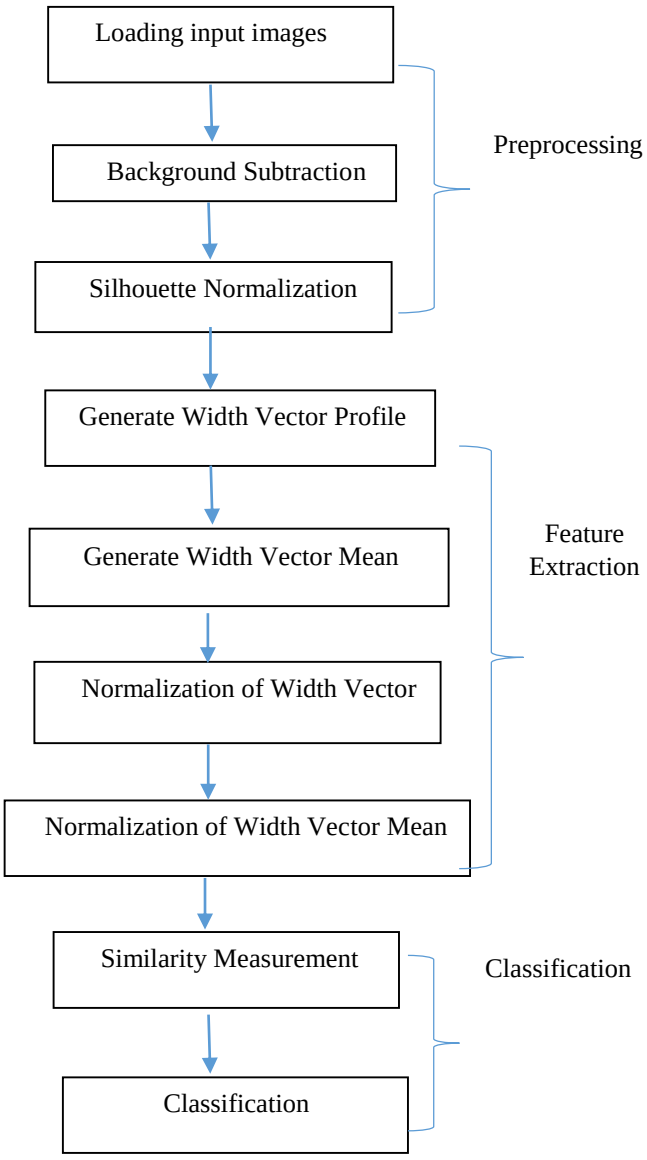


Fig.2.Proposed gait recognition system

II. LITERATURE SURVEY

Liang Wang, proposed a simple gait recognition algorithm using Eigen transformation which is based on Principal Component Analysis. With the PCA time-variant distance signals are generated for sequence of silhouette images. This approach is view dependent [3].

Kale, proposed a HMM-based approach to recognize gait sequence. However this method is robust to changes in speed .But this method not robust to drastic changes in clothing and illumination [4].

Murat, uses distance vectors are projected for binary silhouettes based on difference between bounding box and silhouette. Mahalanobis distance is applied for classification [6].

Soumia Benbakereti, proposed a dynamic time wrapping distance measure which helps to check the similarity of shapes in two time series sequence .He uses dynamic programming technique to get the optimal results. However this approach takes high computational complexity [5].

Mohammad Ali Saghafi, discuss till now the methods available for features extraction taken from images and videos, Similarity measures used for classification their advantages and disadvantages and also discuss some resolved issues remain in person re-identification techniques [7].

In our proposed system, we used boundaries of silhouette images (taking outer contour) which helps to achieve results more accurately. As we are using silhouette of gait which are insensitive to color and texture of cloth .As we used model-free approach which helps to reduce the computational cost. Applying of normalization we reduce the dimensionality of input feature space.

III METHODOLOGY

A.Background elimination

Background subtraction was done on input images. To eliminate background from the image, frame differencing approach is used. In frame differencing we have to take two frames. One is current frame and another one is reference frame.

B. Silhouette Normalization

In our system we take subjects from CASIA database of A dataset. We took the oblique view (45^0) and lateral view (0^0) with respect to the image plane.

Images of silhouette shown in Fig.3.

Fig.3.Images of silhouette

C. Creation of width vector profile

Width vector is the difference between leftmost pixel and rightmost pixel of the outer contour of a silhouette image. Here we create width vector profile for each input gait cycle. The created width vector profile for the input gait cycle as shown in Fig.4 as follows.

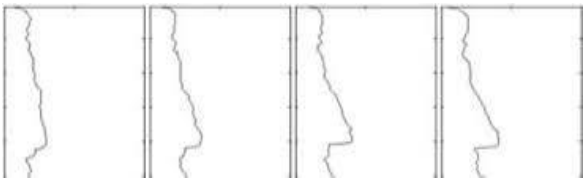


Fig.4.Width vector profile

D. Creation of Width vector mean

To reduce the time complexity instead of using entire silhouette image we are taking width vector mean as a feature. Width vector profile is generated by calculating mean of width vector profile. The created width vector mean for the width vector profile is shown in Fig.5.as follows.

Fig.5.Width vector mean

E. Normalization of features

To obtain results more accurately we normalize the both width vector and width vector mean data. Here we use size normalization. Normalization helps to reduce the effect of size changes when silhouettes are in walking.

F. Similarity measurement

Here we applied different distance metrics to re-identify a person. Experiments done on Euclidean, Manhattan and Canberra distance metrics. Each width vector is compared with the each width vector mean. Among all the distance metrics, Euclidean distance gave us good result for without normalization of feature vectors. Canberra distance metric gave us with normalization of feature vectors. Results are shown for these distance metrics in Section V.

G. Classification

After applying of distance metric we fixed some threshold value. If the distance between width vector of a person and the width vector mean is lesser than the fixed threshold value then we conclude that those two persons are identical. We perform this experiments on CASIA database results shown in Section V.

IV ALGORITHM

We used following steps to perform person re-identification.

Step 1 Load one gait cycle of silhouette images from database as an input.

Step 2 Convert all the input images into .bmp format.

Step 3 Finding x-coordinate values of leftmost nonzero-pixel (X^L_y (t)) and rightmost nonzero- pixel (X^R_y (t)) along same row.

```
/*to select leftmost nonzero pixel*
    Color c=bitmap.GetPixel (i, j);
     If (c.R==255)
       If (i>0)
         Select pixel and store in array;
       End
      End
```

Likewise we have to calculate rightmost nonzero pixel value.

Step 4 Subtract these pixels to generate width vector profile (W_y (t)) as follows.

$$Wy\ (t) = X^R y\ (t) - X^L y\ (t) + 1$$

Step 5 Calculate the mean of width vector mean for the width vector profile as follows.

$$Wy = 1/T \sum_{t=1} Wy\ (t)$$

Step 6 Apply normalization to scale all the width vectors and width vector mean's comes into same range of values.

Step 7 Applying distance metrics between width vector and width vector mean calculated as follows.

Euclidean distance (D (p, q)) = Sqrt $(p-q)^2$

Manhattan distance (D (p, q)) = (p-q)

Canberra distance (D (p, q))

=Abs (p-q)/Abs (p) +Abs (q)

V EXPERIMENTS AND RESULTS

The performance of the proposed algorithm has been evaluated on CASIA database of dataset A. We took lateral view (0^0) and frontal view (45^0) of five different subject's indidually. After that we extract features from the subjects. The correspond width vector profiles generated for each subjects in the gait cycle are shown in Fig.6 .as follows.

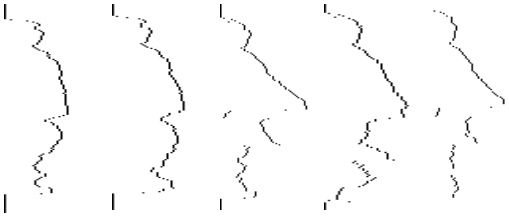


Fig.5.Width vector profile

Width vector mean generated for the above width vector profile shown in Fig.6.as follows.

Fig.6.Width vector mean

After the creation of width vector mean we apply normalization for both width vector profile and width vector mean. For the classification we apply distance metrics (Euclidean, Manhattan, Canberra) between width vector and width vector mean. Without normalization of features results shown in following tables (*A* and *B*).

A. Without normalization ($Th=45^0$)

Person	Euclidean(%of match)	Manhattan(% of match)	Canberra(% of match)
1	100	100	100
2	73	72	50
3	90	90	90
4	100	100	95
5	81	80	93

B. Without normalization ($Th=0^0$)

Person	Euclidean(%of match)	Manhattan(% of match)	Canberra(% of match)
1	86	85	85
2	93	92	92
3	100	98	100
4	100	100	100
5	92	91	91

With normalization of feature vectors results shown in following tables(*C* and *D*).

C. With normalization ($Th=45^0$)

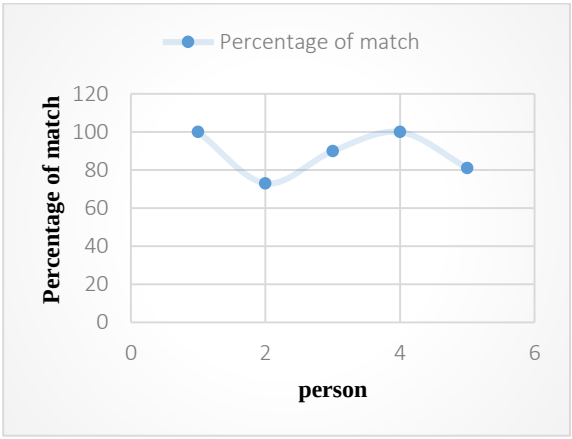
Person	Euclidean(%of match)	Manhattan(% of match)	Canberra(% of match)
1	91	90	72
2	90	90	100
3	80	80	90
4	100	100	90
5	50	45	63

D. With normalization ($Th=0^0$)

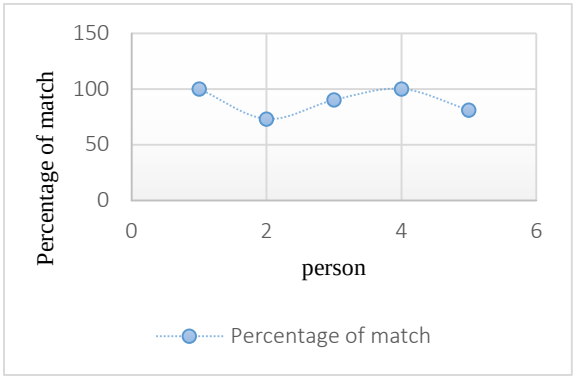
Person	Euclidean(%of match)	Manhattan(% of match)	Canberra(% of match)
1	71	71	71
2	69	69	100
3	86	80	87
4	64	64	71
5	58	50	91

On the above analysis of results, by applying of normalization Euclidean distance gave the best classification rate. With normalization Canberra distance gave the best classification rate.

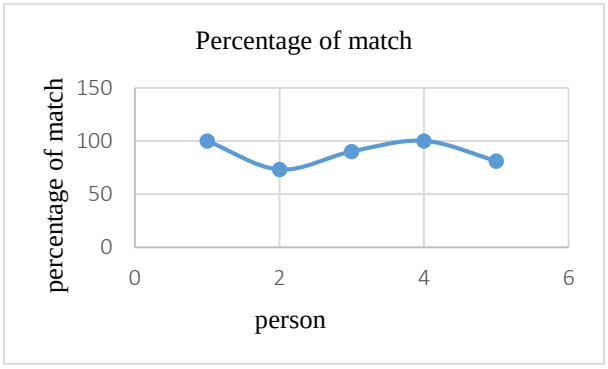
We plot the graph for with normalization (0^0) as follows.



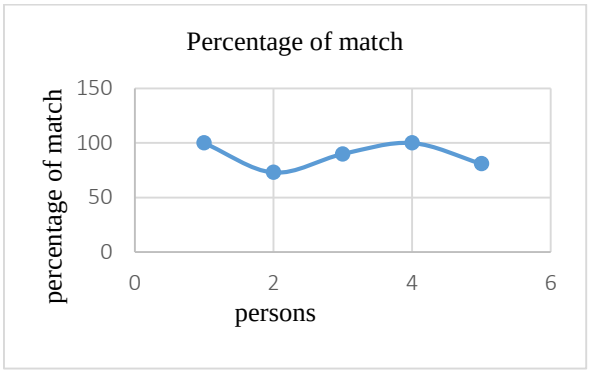
We plot the graph for with normalization (45^0) as follows.



We plot the graph for without normalization (45^0) as follows.



We plot the graph for without normalization (0^0) as follows.



VI CONCLUSION AND FUTURE WORK

In this paper, we present a model-free gait based approach. The extracted features width vector profile and width vector mean is generated for input gait cycle. After that we applied normalization for the extracted features. After that we applied distance metrics between width vector and width vector profile. Compare with the previous existing methods our proposed system gave, without applied of normalization we achieved re-identification rate is 90%. Applied of normalization we achieve re-identification rate is 84%.

The work further extended by providing multi-camera communication software to reduce false counter which gives results more accurately. By including of gait features with some other biometric features like Iris, thigh etc. also helps to increase system performance.

REFERENCES

[1] Jin Wang,Mary She ,Abbas Kouzani and Saeid Nahandi,"A Review of Vision-based Gait Recognition Methods for Human Identification", IEEE,2010.

[2] Angela D'Angelo and Jean-Luc Dugelay, "People re-identification in camera networks based on Probabilistic Color Histograms", Multimedia Communication Department Sophia Antipolis,France,Vol. 7882, 23-27 January, 2011.

[3] Liang Wang, Tieniu Tan ,"Sihouette Analysis-Based Gait Recognition for Human Idenitification",IEEE,2003.

[4] A.Kale,A.N.Rajagopalam,N.Cuntoor and V.Kruger,"Gait-based Recognition of Human using Continuous HMMs",in proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition,(2002),pp.336-341.

[5] Soumia Benbbakereti ,"Recognizing Human Gait in Video Sequences", University of Sciences and Tecnology, IEEE, 2012.

[6] Murat Ekini ,"Human Identification using Gait",Dept.of Computer Engineering ,Vol.14,No.2,2006.

[7] Mohammad Ali Saghafi , Anil Hussain and Halimah Badioze Zaman,"Review of Person re-identification technique" ,IET Comput.Vis.,2014,Vol.,Iss.6,pp. 455-474.

[8] Apurva Bedagkar-Gala and Shishir K.Shah,"Gait-assisted Person Re-identification in Wide Area Surveillance", University of Houston, Dept.of.Computer Science.

A Comprehensive Survey On Big Data Analytics And Techniques

K.Naresh Babu[1]
Asst.prof,Dept of IT ,Geethanjali college of engineering and technology,Hyderabad ,India,
Naresh.kosuri@gmail.com

Dr.Suneetha Manne[2]
Prof and HOD,Dept of IT,Velagapudi Ramakrishna Siddhartha Engineering college,
Vijayawada,India,
Suneethamanne74@gmail.com

Abstract:

Big Data has been developing from few years and created hype. But it is quite normal that 3V's (velocity, volume, and variety) are beyond a more thorough discussion of data approach.'Big Data 'is analogous to small data.Though Big data is Having different approaches, methods, tools and architectures,It need to enhance a better way to solve New problems and old problems.Decision makers must be cautious about customer interactions, daily transactions and social network data,inorder to obtain great worth understanding from various and quickly transforming data. The importance of advanced analytics for big data technologies, can be presented in a stochastic manner. Some of the different analysis techniques which can be a implemented on big data,and the Benefits provided by big data analytics in various areas are been analyzed here.

Keywords:Big data, Big data analytics,Big analytics techniques.

1. INTRODUCTION:

The term Big data is the subject of regard from the municipal planners and academics,corporate leaders, and little extent, there is threat also. The unexpected growth in big data has left many surprised. The swift development of big data technologies, left small time for talk for receipt of the idea by the public and private sectors to enhance and grow-up in the educational area.Clearly, Volume plays a major role while answering about Big data[1]. For example, in the Laney (2001), the volume, variety, and velocity (or three V's) challenges in all three dimensions in the data to propose the three V's data describe big management. "Big data is high-volume[2],velocity and variety information assets that request cost-effective,innovative structures of data processing for improved intuition and decision making.

How big data is defined? The seven V's form fig.1.1 depicts the importance of Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value

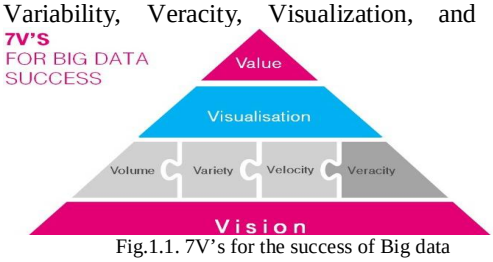


Fig.1.1. 7V's for the success of Big data

Volume

Now, what is measured in Zettabytes (ZB) or Yottabytes(YB) is measured previousely in gigabytes(GB). IOT (Internet of Things) is creating a significant growth in data.

Velocity

Velocity is the pace at which the data is accessed

Variety

Variety is one of the main problems with the big data. It is unstructured, which includes various types of information from video to XML and SMS. Especially with the rapid changes in the data it is not simple task to organize the data in a meaningful way.

Variability

It is different from the types of variability. 6 different blends of coffee is offered in a coffee shop, but you get the same mixture every day, and if it tastes different every day, that is the difference. If you are constantly changing, meaning that it has a huge

impact on your data types, data is true.The accuracy of the data is most important .

Visualization

Visualization is crucial in today's world. Using charts and graphs to visualize large amounts of data and complex spreadsheets and reports are chock-full of numbers and formulas will be more effective than the conveying meaning.

Value

The value of the end of the game. Time, effort and it takes a lot of resources.Addressing the volume, velocity, variety diversity, accuracy, and visualization, then you want to be sure that your company is getting the value from the data.

Today's Challenges

They are available for those who are faced with the challenge of big data, giving marketers. We all have a great appetite for data, but always "digest" is not easy. Data often resides in a different point SOLUTIONS[3]. It is possible to structure the data that make it difficult to merge data from various sources, there are inequalities. The ability to understand the business as a way to gather and present information, so it can make decisions quickly is the key to keeping the competition, the growth of data i.e structured and unstructured will present challenges as well as opportunities.

1.1.Bigdata analytics

Big data analytics is the method of inspecting the hidden designs in large data, unknown associations, market drifts, customer tastes and other big data to uncover insights that can be divided into two major sub-systems(data management and analysis) in the process of extracting useful business information[4]. Fig1.2 depicts the Big data processes. Collection and storage of data management systems and supporting technologies for the analysis of the data and prepare and will return to it. Analysis, can also be referred to the approaches used to obtain, analyze and gain intelligence from large data.

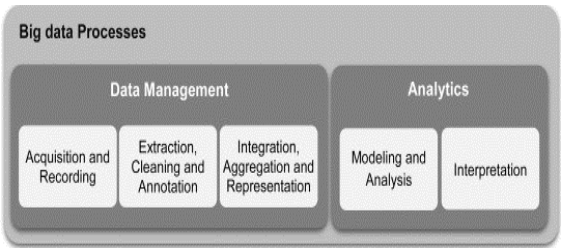


Fig1.2.Big data Processes

Although necessary advances in analytical techniques on large data yet to be taken place in the near future, such as whether the emergence of novel diagnostics. For example,As location-recognition is considered in understanding social media,It is a constructive field of research. Big data is highly interrelated, noisy and is unreliable,So it leads to the development of statistical methods[5] to large data mining to everybody, while rest delicate to the unique characteristics.The more effective marketing, better customer service, leading to new revenue opportunities, rival companies and other business interests the operational efficiency, improved.competitive advantages.

Literature survey:

1. John A. Keane, in 2013, which can be a model for the development of big data applications Shape three stages (Data from multiple sources, data analytics and designing, data administration and exposition) and seven layers (presentation layer,access layer, modeling, processing layer, system layer, data layer, multi-sample layer)to divide the data in the application. The main purpose of this paper is a huge amount of big data applications is to manage and architect. The purpose of this paper and the draft Bridge, big data, high-performance business requirements, the technology gap with the realities of diverse data and data sources in a timely manner is carried out. The issues of this paper is very difficult to combine with present data and systems.

2. Dong Jin Luna In 2013, the Data combination (schema mapping, record linkage and data binding) explained the challenges. Examples of big data to address the challenges raised by these new challenges and data integrity using the methods explained, the volume and sources, speed, and accuracy of the number. The purpose of this paper is to identify

problems with the data source to combine with current data and systems. The complication of this paper, such as crowd sourcing markets, integrating data exploration tool for data sources providing data integration is the integration of large data.

3. Jun Wang In 2013, the performance, capacity, and delays in the implementation of the data to improve issues such as Data-grouping aware of the (drawn) data placement on the proposed scheme. Hadoop map reduce compared to the small number of nodes that could be a cluster of several groups of data. There are three main steps to draw as defined in this paper: the data by grouping information from system logs and data in one place to learn to recognize, the cluster grouping the data matrix. The advantage of paper, to improve the throughput of up to 59.8%, up 41.7%, to reduce the execution time and Hadoop / map is to improve the performance and reduce the total by more than 36.4%.

4. Yaxiong Zhao In 2014, aware of the data caching (Dcache) framework map using the model to reduce the increase in the processing of large data applications programming model to reduce the minimum number of the original map of the proposed change. Description of the scheme and structure of the data cache is known as a protocol. The advantage of this paper is to reduce the full-time jobs in the map improves.

2.Bigdata analytics: Techniques for big data analytics

There are many methods being used to analyze datasets[6].This paper provides a list of some techniques applicable. This result is by no means confined, researchers continue to develop new techniques to enhance existing ones, particularly in need to analyze new combinations of data.

2.1.Text analytics

Text Analytics is used to convert unstructured data to relevant data in order to calculate customer mind, product evaluation and feedback,it also provides search provision, sentimental analysis and entity modelling,in order to support reality based decision making. Text analysis uses much linguistic, statistical, and machine learning techniques. Text Analytics make use of information retrieval from unstructured data and to obtain patterns,drifts and measuring and interpreting the output data from the process of arranging the input text.The techniques like Lexical analysis, categorization, clustering, designs recognition, tagging, information mining,visualization, and predictive analytics are also involved in this process. Text Analytics includes key words, concepts, classifications,meanings, tags from different sources of text data are available in many files and formats[7]. The results, which are extracted entities, realities, connections are commonly stored in a relational, XML, data warehousing applications and are analyzed by tool like business intelligence, big data analytics or predictive analytics.

2.1.1. Process Flow of Text Analytics

Text Analytics system process :

- Text: Data is unstructured in initial stage.
- Text organizing: The Data will be shifted in Semantic text.
- Text modification:Main text will be extracted for further use.
- Feature choosing:Data is measured and displayed in Statistical manner.
- Data mining:Data here is clustered and classified.

2.1.2.Features of Text Analytics

- Extraction of ideas, entities, relations, events.
- Search entry,indexing,equivalent document identification.
- All vital file formats are examined.
- Link investigation, link text database
- Able to detect and examine sentiments.
- Document summarization and records management.
- Interactive presentation.

2.1.3.Applications of Text Analytics

- Sentiment Analysis.
- Find entrance of unstructured data.
- Placement of adds automatically.
- Monitoring Social media.
- Data mining and Business intelligence.
- Records Organization.
- National security and intelligence.
- discovery-based science

2.2. Audio analytics:

Audio Analytics means the extraction of meaning and information from audio signals for Analysis.Audio Analytics is added to give more attention to stay alert in conditions where keeping an overview is hard[10],also makes it possible to estimate situation in which video analytics cannot do by themselves.Audio analytics is worth full and can be easily extended upto present video surveillance. All Most all (IP) cameras are equipped with audio input or microphone.Analyzing of ambient audio can start when the required software is installed on the camera or a connection is established between the camera and computer with the audio analytics software installed. Infact, advanced audio analytics will be used when video surveillance fails.

What we hear is a group of patterns perfectly analyzed and accepted by our brains. Advanced Audio Analytics mimics the process of human hearing and allows sounds by using advanced algorithms. This is advanced beyond traditional sound recognition[12],based on volume and time thresholds. Advanced Audio Analytics identifies specific sounds, such as aggression, breaking glass etc. if they are combined with ambient noises,Even at a low volume.Advanced Audio Analytics decreases the need to record video streams and thus reduces possible violation of privacy.

A method is proposed for automatically recognizing observed audio data is shown in fig.2.2. An observation vector(OV) is created of audio features extracted from the observed audio data and the observed audio data is recognized from the OV.

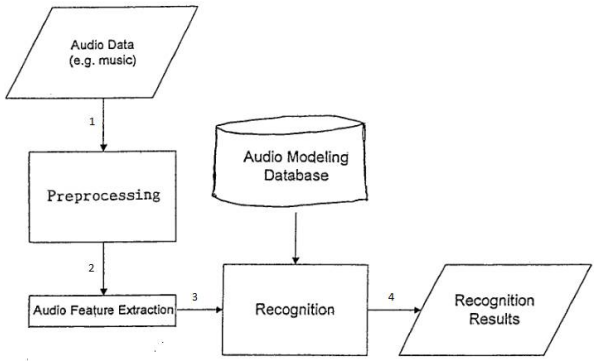


Fig.2.2.Audio analytics recognition process.

2.2.1.Features of audio analytics

- Response is faster
- Potentially serious problems are eliminated
- Incidents are Less missed
- Employee gets alerts automatically
- Many clients exists per employee
- Safety feeling is expanded

2.2.3.Audio Analytics Applications:

Audio is termed as a file format that is help full to transmit data from one place to another. Audio analytics checks whether given audio data is available in desired format that sender sent. There are man audio Analytics Applications,few are mentioned below:

a)Supervision: Supervision is based on process of organized option of audio category in finding crimes happening in society. The Supervision is based on audio Analytics structure,It is the only method to identify doubtfull activity. The application is used to send main information to supervision at some problematic situations immediately.

b) Threats Detection: Audio is help full to detect the threat taking place between sender and receiver.

c)System of Tele-monitoring: Present technology uses camera to record audio. Audio Analytics facilitates detection of loud voices,glass breaking, sound of gun,explosions etc. Combination of audio Analytics and video Analytics result as a best threat detection efficiency.

d)Mobile Networking: Mobile networking is used to transmit data from one place to another place.Due to network problem sometimes audio doesn't work exactly at that particular time Audio Analytics is used to detect the information which is not sent properly.

2.3Video analytics:

Video analytics[13] is the process of mining structured data from unstructured data.Many of us have been involved with video analytics from many years.Video motion identification is the basic method of video analytics.Currently sophisticated video analytics is making us to see more, watch less, yet know much.The costly false alarms are minimized considerably.Threats are detected much earlier, heading off potential lawsuits. There are several key contributors to the enhancement of video

analytics:The requirement, technology,network and support.The process of analyzing of surveillance footages, the gathering of information from them and connecting a crime and the perpetrators is known by different other names.Video analytics is analogous to 'video content analysis' or 'computerized video analytics' in the course of any research.

2.3.1.Video Analytics Architecture:

Video Analytics is been implemented in three configurations.

EDGE-BASED SYSTEM

Network camera performs analyses of image and gives an alarm signal to the operators based on pre-configured changing rules.

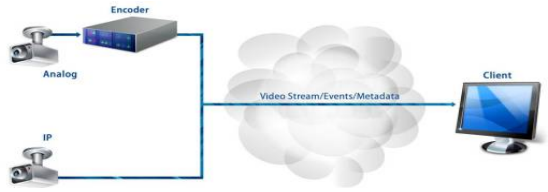


Fig.2.3.1.Edgebased system

Edge-Based System doesn't require high-performance central server and it makes the application more scalable, reliable and cost-effective.

SERVER-BASED SYSTEM

Server-based system authorize more complex analyses. All the images captured by cameras are transferred to the central server and the server analyzes

Fig.2.3.2 server based system

them with the processing power, wide memory space, higher-speed data base access and more advanced software.

HYBRID SYSTEM

Hybrid system is a combination of edge-based system with server-based system and significantly reduces the overload of server and network.Smaller system runs Intelligent Video applications. Suppose a system which detects a person from visitors.Every captured image with photos on data base are compared.

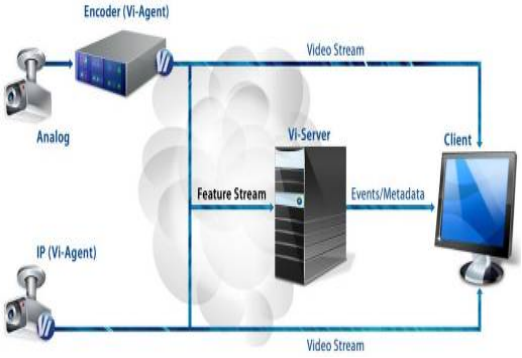


Fig2.3.3 Hydrid system

The server actually needs the facial part of captured image. The rest wastes the resources of server and network. Hybrid system optimizes it. Cameras clip the facial part on the edge and the server does a comparison.

2.3.2.Features of Video Analytics

- Security is enhanced
- Operational Cost is Saved
- Increase in Performance
- Safety and Compliance are maintained
- Service quality improvement.

2.3.3.Applications

stationary objects motion: For example,The operator will receive an alarm when an exhibit in a museum starts moving.

Detection of objects with No motion: For example no motion of deserted luggage or vehicles in sensitive areas

Monitoring of Virtual Tripwire : Used in detecting trespasssers in high security sites like Airport Runways, Railway Tracks, etc.

Motion Direction of Objects: A flow moving in the opposite direction to another can be identified where movment is done in either a wrong or unpredicted direction such as the wrong way in a one way system and also to identify people, vehicles etc. visiting sites under supervision

Pottering Detection: Pottering can be a predecessor to crime,a person or persons moving around a site particularly a known trouble spot will be identified.

Counting of Peoples or Vehicles: It is used to identify congested situations in queues at railway stations, super markets etc.To identify the number of people visiting different parts of a shopping centre is use full in marketing purposes so that rents can be calculated.

2.4.Social media analytics

Social media analytics explains the process of collecting and combining raw data from social media like Facebook, Twitter, and blogs and even forums like customer support and user communities and analyzing them to support planning and decision making. As many people around the world use social media sites to communicate with their friends and family members, social media analytics is playing a vital role in branding, customer acquisition and retention, and other sales and marketing strategies.

2.4.1.Social media analytics process

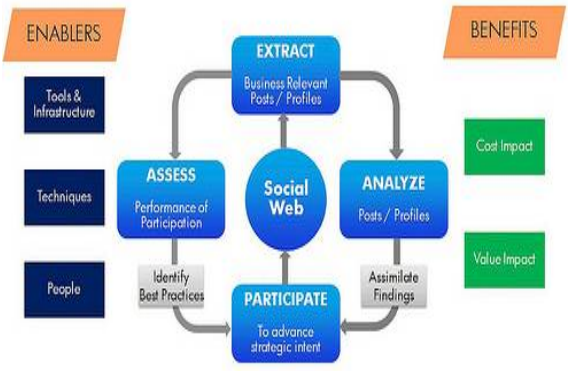


Fig.2.4.Social media analytical process

Social media analytical process is depicted in Fig.2.4.Initial step in social media analytics process is extracting business relevant data.Data extraction can have 2 different scopes. For requirements such as campaign monitoring, the scope is all posts from entire social media universe that match to a defined set of keywords or search terms. On the other hand, for requirements such as performance measurement or competitive intelligence, the scope is all posts from a defined set of social media profiles.

In ANALYZE step we try to clean and make sense of the gathered data.Aspects such as volume trend analysis, ranking posts, ranking profiles, etc may be involved. EXTRACT and ANALYZE steps performed on a regular basis comprise a social media listening program. The discovery and insights from a listening exercise could feed into various business purposess like development of product, customer support, sales, etc. as mentioned in the social analytics life cyle defined by Ken Burbary and Chuck Hemann.

Discoveries from listening exercise should also provide inputs into a brand's social media participation strategy and plans (e.g. as simple as identifying popular topics relevant to a business in order to become part of those conversations).If there is active participation, it naturally demands performance assessment for continuous improvement. Focus should be on identifying best practices from the participation experience. This completes the ASSESS part of the cycle.

2.4.2.Features of social media analytics

- Competitive Advantage can be gained
- Learning from Your Customers
- Your Products and Services can be enhanced
- Better Target Marketing Efforts
- Market Innovation.

2.5. Predictive analytics:

Extracting information from existing data sets to determine patterns and predict future outcomes and trends is termed as predictive analysis.Analyzing current and historical facts to make predictions about the future is depicted in Fig.2.5.

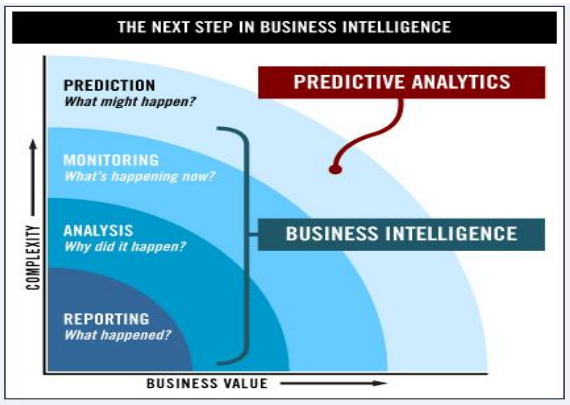


Fig.2.5 predictive analysis

2.5.1.Features of Predictive Analytics

- Models can be used to predict customer attributes and behaviour.
- Using probability, models can anicipate outcomes.
- Very Powerful and Profitable.

2.5.2.Applications:

- Direct Marketing
- Customer Retention
- Fraud Detection
- Risk Management
- Clinical Decision Support Systems.

3.Conclusion and future work:

Big data is an innovative topic,which is comprehensively examined here, which has recently gained interest based on its realistic unique chances,advantages.Present day world is currently living in, many varieties of high velocity data produced daily, and lay intrinsic details and patterns a within them of concealed knowledge to be mined and utilized.Thus, big data analytics can be applied to hold business change and increase decision making, by applying advanced analytic techniques on big data, and revealing hidden insights and valuable knowledge.There are many challenges for future research with Big Data,a few are addressed below.

Challenges with Big Data:

- Technical Challenges:
- Capturing & Analysis of data
- Storage & Sharing/Transferring
- Searching and Visualizing

4.References:

[1] V. Mayer-Sch¨onberger and K. Cukier, Big Data: A Revolution that Will Transform how We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt, 2013.
[2]http://www.smartdatacollective.com/michelenemschoff/206391/quick-guide-structured-and-unstructured-data.

[3]. D. S. Modha, R. Ananthanarayanan, S. K. Esser, A. Ndirango, A. J. Sherbondy, and R. Singh, “Cognitive computing,” Communications of the ACM,vol. 54, no. 8, pp. 62–71, 2011.

[4]. www.microsoft.com/bigdata.

[5]Peter Harrington, Machine learning in Action, Manning Publications, 2012.

[6]Van der Valk, T., Gijsbers, G.: The Use of Social Network Analysis in Innovation Studies:
Mapping Actors and Technologies. Innovation: Management, Policy & Practice 12(1),5–17 (2010)
[7]. Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing
Systems (ICDCS), pp. 25–36 (2011)
[8]. Panasonic_Intelligent-Video-Technology_Whitepaper
[9]. A. Kumar, A. Beutel, Q. Ho, and E. P. Xing. Fugue:
Slow-worker-agnostic distributed learning for big models on big data. In AISTATS, 2014.
[10]Big Data, for better or worse, http://www.sciencedaily.com/releases/2013/05/130522085217.htm , Accessed on 3 Jun, 2015.

[11]. Hortonworks. http://hortonworks.com/.

[12]Demchenko Y, Grosso P, de Laat C, Membrey P. Addressing big data issues in scientific data infrastructure. In: 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, 2013. IEEE, pp 48–55.learning. In: ACM Sigmetrics: Big Data Analytics Workshop. ACM, Pittsburgh, PA
[13]. Bogdan Ghit¸, Alexandru Iosup and Dick Epema
“Towards an Optimized Big Data Processing System”,
13th IEEE/ACM International Symposium on Cluster,
Cloud, and Grid Computing, pp. 83-86, 2013.
[14] Hirudkar AM, Sherekar SS (2013) Comparative analysis of data mining tools and techniques for evaluating performance of database system. Int J Comput Sci Appl 6(2):232–237
[15] D. Stauffer and A. Aharony, Introduction to percolation theory. CRC press, 1994.
[16] J.-P. Onnela, J. Saram¨aki, J. Hyv¨onen, G. Szab´o, D. Lazer, K. Kaski, J. Kert´esz, and A.-L. Barab´asi, “Structure and tie strengths in mobile communication
networks,” Proceedings of the National Academy of Sciences, vol. 104, no. 18, pp. 7332–7336, 2007.
[17] V. Frias-Martinez, C. Soguero, and E. Frias-Martinez, “Estimation of urban.

AUTHOR’S PROFILES

Dr.Suneetha Manne is a professor and Hod, Dept of IT at Velagapudi Ramakrishna Siddhartha Engineering college,Vijayawada,India,She completed her P.Hd from Osmania university,Hyderabad,India.Her research lies at the intersection of Data mining and IT.She is specifically focused on big data analytics it relates to Text analytics.

K.Naresh Babu is an assistant professor at Geethanjali college of engineering and Technology,Hyderabad,India.He is pursuing his P.hd in Data mining from JNTU,Kakinada and his research includes Big Data Analytics.

GIS the future of Utility Management

P Sree Gayathri [1], V Phani Kumar [2]

[1]*PG Scholar, Dept. of CSE, V.R Siddhartha Engineering College,* [1]E-mail:psreegayathri@gmail.com

[1] *Asst Professor, Dept. of CSE, V.R Siddhartha Engineering College,* [1]Email: phanikumar.venna@gmail.com

***Abstract*—Energy conservation is one of the burning issues now-a-days due to tremendous scarcity of electricity across the country. As a consequence the power sector role and overall growth of economy is important and critical. But transmitting electricity over distance and via networks involves energy losses. To reduce these power losses throughout the country and use the electricity in an efficient way, GIS can be used in possible applications to determine optimal path for transmission lines and locate fault transmissions. The broad scope of this study is to provide the information system on electrical assets for the electricity board users through Geographic Information Systems [GIS].This will be made available on query for tracking any asset in the entire network of a selected town or towns. Users can search the substation by its name, capacity, type, section name, etc., and also the system allows user to search HT, LT Lines, DTR's and Poles based on the selected fields and values. System will also generate the reports for the substation, DTR, and Substation Elements as per the user choice. In this system the search is basically a non-graphic search and the final result will be displayed on a GIS map.**

Index Terms:- GIS, Utilities, HT, LT lines, DTR, non-graphic

I. INTRODUCTION

The power sector constitutes an important aspect of utility domain and the backbone of the national economy for any country in the world. Adequate electrical power with a high degree of quality and reliability is also the key to Indian economic growth. India is the 3rd largest producer of electricity in the world. Regardless of this growth in the generating electricity India is facing huge power deficit. In this overall development of power sector in India transmission and distribution system constitutes crucial link between the generating and consumption sources. However this distribution system has grown in an unplanned manner to meet the growing demands of consumers on an urgent basis which in turn contributed to very high Aggregate Technical and Commercial losses (AT&C) along with poor quality and low reliability of power supply to the consumers. The most challenging factor for many utility companies is to maintain, manage, model and map their distributed facilities and networks optimally to meet the customer expectations and industry compliance regulations.

As the electrical utility networks continue to grow in complexity and size, the probability of two or more networks occupying a common right-of-way or intersecting each other keeps on expanding. Due to this there are some conflicts in fault management faced by the utility management sectors of the country. The existing ones are very time consuming and man power consuming which tends to loss due to frequent thefts and also insufficient supply of the electricity. By developing GIS based web mapping applications the electricity board users can effectively visualize and analyze the conflicts in the electrical network and resolve them by reducing the power loses.

II. LITERATURE SURVEY

A. Existing Electricity Distribution System in India:

Power distribution is the most crucial link in the electricity supply chain and, unfortunately, the weakest one in the country segment as it has a direct impact on the sector's commercial viability and ultimately on the consumers who pay for power services. A big challenging factor is transmission and distribution (T&D) losses which are estimated nearly up to 30% overall. There are many technical and non-technical factors that are contributing to high T&D losses which were computed taking into account electricity bills issued to consumers as accrued income and not on the actual collection. The concept of Aggregate Technical & Commercial (AT&C) losses has been introduced in 2001-2002 to capture the difference between the billing and collection, which was not captured by the T&D loss figures. AT&C loss is given by the difference between units input into the system and the units for which the payment is collected.

B. Reasons for AT&C losses:

There are mainly two categories of reasons for these AT&C losses:

- Technical losses
- Commercial losses

The Technical losses are due to following reasons:

- Overloading of existing lines and substation equipment's.
- Absence of up gradation of old lines and equipment's.
- Low HT:LT ratio
- Poor repair and maintenance of equipment's
- Non- installation of sufficient capacitors.

The Commercial losses are due to following reasons:

- Metering Inaccuracies and error in meter reading.
- Unmetered losses of very small load

- Absence of Energy accounting and auditing
- Billing Problems

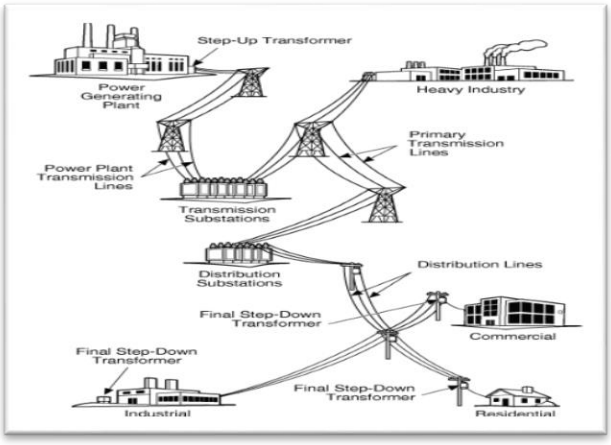


Fig.1. Power Distribution Network

Today power distribution is maintained by manual updating of planned network construction operation and also lack of monitoring, less prevention of losses and manual update of consumer records etc., Hence the electric distribution are realizing the benefits of GIS Geographical Information System that are designed for electrical utilities which tends to manage the distribution system to reduce the power outages providing a geographically oriented view of electric distribution structures.

C. Advantage of GIS in Electricity Distribution network:

GIS is a computer-based platform capable of handling spatial data that represent real world feature transmission lines, in the form of digital maps and attributed geo-relational database. This model allows new methods to be used and provides high-quality presentation of processed data and decision making tool in situations when data relevant to a decision include a spatial component.

Now- a-days utility sectors are realizing the benefits of using GIS technology in the area of automated mapping. Consequently GIS has become a very significant tool for electrical utilities and for activities like utility asset management, maintenance planning, outage management and network planning, especially in the area of distribution planning. These details should be accurately maintained and make sure up-to-date as possible in GIS which help in providing necessary information for building an electrical network which ensures the growth and reliability in the network.

D. Survey through GIS and GPS

GPS: Global Positioning System (GPS) is a system composed of a network of 24(now 29 satellites) well-spaced satellites that orbit the earth and make it possible for people with ground receivers to pinpoint their geographic location. For tracking the data on the ground, GPS receivers and rovers are used. The ground control points (GCP's) are collected which represents the real-world pseudo coordinates to be further used for geo-referencing the satellite data as depicted in the following image

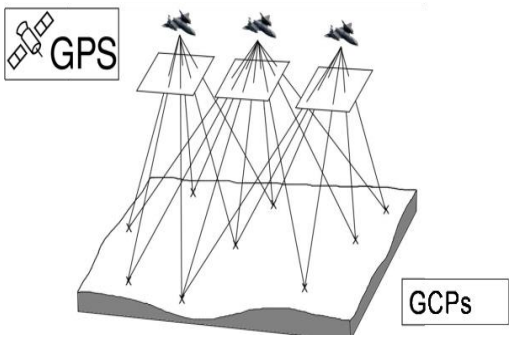


Fig.2. Recording Ground Control Points

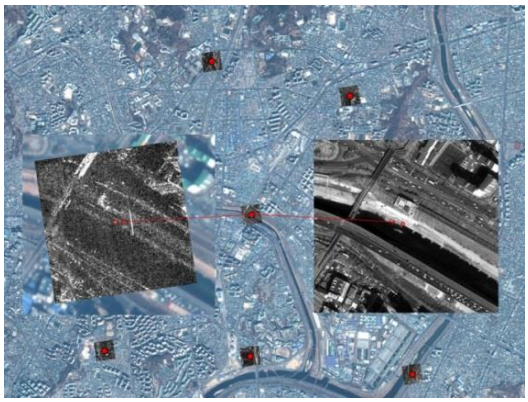


Fig.3. Satellite Image for GCP's

D. How DGPS works?

Differential correction techniques are used to enhance the quality of location data gathered using global positioning system (GPS) receivers. The differential correction can be applied in real-time directly in the field. The underlying premise of differential GPS (DGPS) requires that a GPS receiver, known as the base station, be set up on a precisely known location. The base station receiver calculates its position based on satellite signals and compares this location to the known location. The difference is applied to the GPS data recorded by the roving GPS receiver.

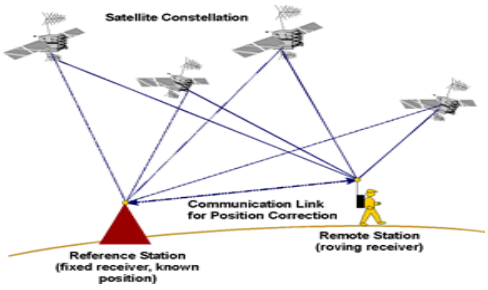


Fig.4. Recording Differential GPS

E. How GPS is used for GIS

GPS is an excellent data tracking tool for creating and maintaining a GIS. It provides accurate positions for point, line, and polygon features. By verifying the location of previously recorded sites, GPS can be used for inspecting, maintaining, and updating GIS data.

Thus the GIS technology has the facility to plan an electrical transformer or distribution network which sometimes is very difficult to access for surveying in the remote hilly areas. Earlier it was done manually on road maps now among modern technologies GIS will be helpful to solve these problems.GIS provides concise platform for system representation and manipulation on network models and data base can be accessed and modified to perform system analysis in an efficient manner.

GIS operates as follows:

1. They accept geographic input in the form of scanned-in and digitized map images. Often this data is supplied by a source that may own maps and has already digitized them.
2. They rescale or otherwise manipulate geographic data for different purposes.
3. They include a database manager, usually a relational database management system (RDBMS).
4. They include query and analysis programs so that you can retrieve answers to simple questions such as the distance between two points on a map or more complicated questions that require analysis, such as determining the traffic pattern at a given intersection.
5. Usually answers provided by them are visual outputs as maps or graphs.

III. Practical Implementation

In many complex web applications which are classic desktop type are being supplanted by rich internet applications [RIA]. These RIA applications basically run on server and the end client's access them through their web browser which intern gives the same experience as in desktop framework. Similarly when common users of current information system come across the word "Geographic Information", most of them are associating this with Google maps, which are web based GIS that has been launched in 2006 by Google. It is an undeniable thought because the simple features of GIS are usually realized by Google Maps.

To analyze the spatial data on the map, to locate the positions and examine the right data we need to integrate applications with GIS features. For delivering such feasible and reliable application to the end-user, development of web mapping application is mainly implemented using ARCGIS Runtime SDK for JAVA which supports best cross-platform development either on Windows or Linux. The applications can be developed with GIS capabilities using the JAVA Web ADF [Application Developer Framework] provided by ArcGIS. The JAVA Web ADF is an AJAX enabled framework for building web applications built on Java EE and JSF which includes server side controls and client side behavior that are accessible via Java Script Libraries.

IV. Implementation Process

How ArcGIS Tool is used:

The ArcGIS tools include ArcGIS Server and ArcMap. By using this tool we can publish the maps as web services so that those can be used by the developers and make them available for the users. GIS will organize the world into layers of geographical objects. These layers contain features modeled as points, lines and polygons and also they have attributes nothing but the information related to them. So the layers are collected and made into a single GIS map using this tool.

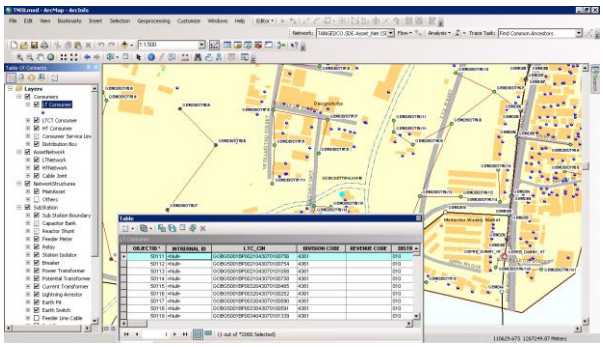


Fig.5. ArcMap image showing the map layers

How JSF and Java are used:

Using the web ADF provided by ESRI ArcGIS those components can be used for creating the web based mapping and analysis applications. These components provide the functionality out of box by allowing easy extension by working with multiple data sources. Hence by utilizing the JSF framework from ESRI we can develop an interactive web based GIS applications. We include richfaces a rich component library tag also into the JavaServer Faces built on an advanced open source framework (Ajax4jsf) which allows easy integration of AJAX capabilities into the application development. For including those features we include the following tags:

- *<%@ taglib uri="http://java.sun.com/jsf/core" prefix="f" %>*
- *<%@taglib uri="http://www.esri.com/adf/web" prefix="a"%>*
- *<%@taglib uri="http://java.sun.com/jstl/core" prefix="c"%>*
- *<%@ taglib uri="http://richfaces.org/a4j" prefix="a4j"%>*
- *<%@ taglib uri="http://richfaces.org/rich" prefix="rich"%>*

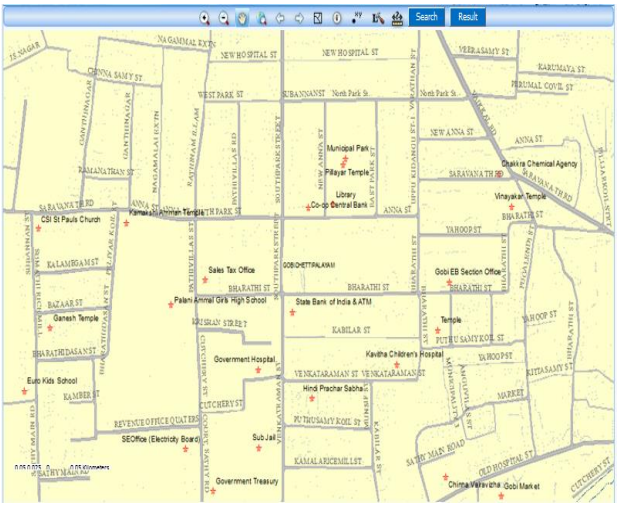


Fig.6. Sample sketch for network shown on Map.

Figure.7. Sample sketch for feeder lines

V. Conclusion

Implementing utility management system using GIS will be a powerful tool for predicting and managing risk factors for any type of utilities. The GIS applications extend over various assorted fields, but it is computer science which holds the key to understand and make advancements of the underlying spatial database and programming for custom applications. To build robust and scalable enterprise applications the JSF JavaEE platform is a rich framework. The applications finally developed will facilitate online query with geographical display, by showing particular assets and their attributes related to utilities.

VI. References

[1] N. Kumar, M. Kumar and S. srivastava, "Geospatial Path optimization for Hospital: a case study of Allahabad city, Uttar Pradesh," 1st ed. IJMER ISSN: 2249–6645, 2014.
[2] N. Parkpoom, "GIS-based model for implementation on Power Transformer planning within Thailand Power Network," 1st ed. China: CIGRE-AORC, 2013
[3] Rao, M.V.K.; Varma, B.S.; RadhaKrishna, C., "Experiences on implementation of GIS based tools for analysis, planning and design of distribution systems," in Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE , vol., no., pp.1-8, 20-24 July 2008.
[4] Dnb.co.in,"India'sEnergySector", 2016.[Online].Available:
[5] https://www.dnb.co.in/IndiasEnergySector/Power_Dist.asp.
[6] E. Maina, "Application Of Gis In Electric Utility Company", 2016. [Online]. Available: http://www.esriea.co.ke/proceedings/GIS%20in%20Kenya%20Power.pdf.
[7] "GIS for Electric Distribution,"2016. [Online]. Available: https://www.esri.com/~/media/Files/Pdfs/library/brochures/pdfs gis-for-electric distribution.pdf.
[8] Esri.com, "Differential GPS Explained". [Online].Available: http://www.esri.com/news/arcuser/0103/differential1of2.html

An Algorithm for Basic IoT Architecture

Rajesh Vemulakonda [#1], Sowjanya Meka [*2], Venkatesh Ketha [#3], Phani Praveen Surapaneni [#4], Sri Vijaya Kondapalli [*5]

[#] *Computer Science & Engineering Department, Prasad V. Potluri Siddhartha Institute of Technology Vijayawada, India*

[1] vrajesh@pvpsiddhartha.ac.in

[*] *Department of information Technology, Prasad V. Potluri Siddhartha Institute of Technology Vijayawada, India*

[2] Sowjanya.meka6@gmail.com

***Abstract*—IoT is the emerging technology empowered by latest developments in various types of Internetworks and Protocols, Communication Technologies, smart devices and sensors, etc… In the first phase of IoT, human interaction is subsided and most of the work carried out by the smart devices. Future technologies in IoT will make more intelligent decisions by using smart sensors enabled physical objects. In this paper, we proposed an algorithm for basic IoT architecture. A brief introduction of IoT along with interactive eco-system and the working nature of algorithm are enlightened.**

I. INTRODUCTION

The technology "Internet of Things (IoT)" was proposed more than two decades ago by researchers from IT industry but came into existence recently. It is a combination of several things, where the "things" in it are nothing but the applications, each application has some specific characteristics to make this technology work in an efficient manner to carry out the specific work to reorder many aspects of the way we live [1].

For client, IoT products such as appliances those are Internet-enabled, media, environmental monitoring, Infrastructure management, Energy management, Manufacturing, Building & home automation and Medical & healthcare systems are moving us toward a vision of the "smart world'', offering more reliability, flexibility, privacy, security and optimal energy consumption aspects [2].

The concept of "smart cities" has less obstruction and optimal power consumption through smart vehicles, smart traffic system, and smart roads. Smart vehicles are well networked vehicles that move around by using geo-graphical networks. Smart traffic signals deviates the traffic by using intelligent algorithms. Smart roads are the roads that are embedded with rich featured sensors [3].

Sensors play a vital role in Internet of Things [4]. Through the collected information from sensors, fields like agriculture, industry, and energy production and distribution increasing the availability along the value chain of production. However, IoT elevate many problems and challenges that need to be considered and approach in order for potential benefits to be realized [5].

II. HUMAN SENSES

In sensing organisms, sensory cells of a system respond to a specific physical phenomenon and maps particular areas in the brain. Human sensing organisms are mainly divided into two types' extroceptive and introceptive [6]. Former deals with the position, motion and state of the body, later with perceive sensations in internal organs. The eyes for sight, nose for smell, skin to touch, tongue for taste and ears for hearing are the five sense organs, first classified by Aristotle. Yet we have several other sensing organisms like receptors in the muscles, tendons, joints, vestibular organs, circulatory systems, digestive system, and etc. The information from this organs send to brain in order to provide the knowledge of outside environment depends on our ways of awareness [7]. In some cases one or more organs might not work properly, in such cases, other properly working organs need to exceed their normal functionality to supply make up information. Perception psychology, Cognitive psychology, and Neurosciences are the specific areas that explain the operation, classification and overlapping theory of senses. To get a quality and comfortable life these human senses are provided in the form of sensors in IoT [8].

III. INTERNET OF THINGS (IOT)

In this section we will have the definition for IoT, basic architecture of IoT, and etc.

A. About IoT

IoT is a rich collection of smart things; the things are from our daily life in order to make our life more comfortable [9]. Smart things are from the substantial collections of intelligent algorithms & applications, powerful sensors, flexible mechanical parts, communication devices, well organized internet, and etc.

As shown in the following Fig. 1, the basic architecture of the IoT contains four levels i.e. sensing devices at the first level, in the second level we have gateways, third level consists of cloud services and the in the final level user

interactive devices are there. The following is the block diagram of the basic architecture of IoT.

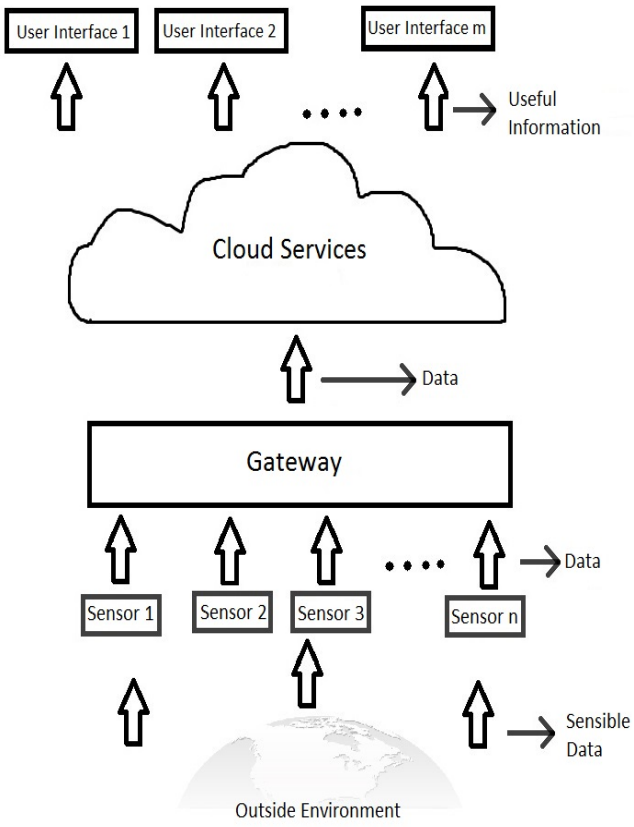


Fig. 1. Block diagram of Basic IoT Architecture.

B. *Sensing Devices*

A Sensor is a minute device which is used to repeatedly perceive and respond to any signal which can be read from physical stimuli, displayed or stored in the form of digital signal. It is used to quantify characteristic of any object. As described in the section 2 about human sensing organisms, in order to make any intelligent decisions, sensors in IoT plays a major role [10].

In this drastic changing technological world wide variety of sensors came into existence. Sensors play a crucial role in Telecommunications, Health care, Business, Industry, Aircraft, Automobiles, Consumer electronics and etc., [11].

Several sensors are categorized according to their nature of work. Automotive sensors (e.g.: Speedometer, Radar gun, Speedometer, fuel ratio meter.), Chemical Sensors (e.g.: Ph sensor, Sensors to detect presences of different gases or liquids.), Electric and Magnetic Sensors

(e.g.: Galvanometer, Hall sensor which measures flux density, Metal detector.), Environmental Sensors (e.g.: Rain gauge, snow gauge, moisture sensor.), Optical Sensors(e.g.: Photo diode, Photo transistor, Wave front sensor.), Mechanical Sensors (e.g.: Strain Gauge, Potential meter (measures displacement)), Thermal and Temperature sensors.(e.g.: Calorimeter, Thermocouple, Thermistor, Gardon gauge.), Proximity and Presences sensors (. e.g.: Doppler radar, Motion detector.) are few categories of sensors. The sensory nature of smart mobiles makes it to play a key role in IoT communication [12].

The following Fig. 2 shows the Block diagram of a SENSOR which contains four major components [13]. The Transducer receiver senses the physical stimuli and passes the information to Logic Circuit which performs necessary action according to the nature of the sensor, the output of Logic circuit is carried out by Detector circuit to convert them as digital signals and the formatted digital signal supplied to the communication devices through the Output module.

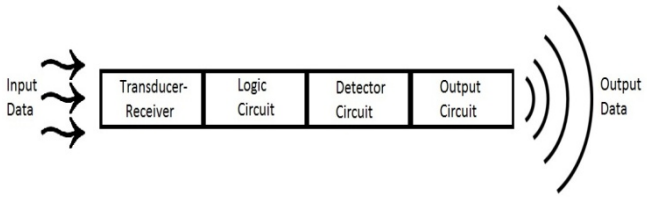


Fig. 2. Block diagram of a sensor

In IoT system, software is like a brain and sensors act as a nerve system, which collects continuous information from the outside environment to get processed by IoT software. The main challenges faced by sensor technology in IoT are sensors need to collect dynamic information which is rapidly changing with reasonable security and reliability, it needs to be cost effective, it needs to show variations as per the real world changes, it has to provide a wide variety of information according to the processing purpose, it should be weather and temperature effective, it has to cover wide range of area, it should elaborated rather than specific in nature, and etc [14].

C. *Gateways*

An application called Gateway neither the server nor the client can access directly, through which client, sever communicates with each other to exchange the data/information in a network [15]. The information from outside of the internet can be securely accessed by the server irrelevant to data/resource, the Gateways make it possible. The main purpose of Gateways is to connect non-IP devices to the internet or network. Gateways can handle traffic from multiple sources [16].

ln IoT, Gateway establishes a communication path between the field and a cloud for processing and storing of data in both online and offline mode. Generally two kinds of Gateways are in use, simple Gateways and Embedded Control Gateways [17]. Where, former organizes and transports the data from/to end points and later extends it functionality by applying intelligent algorithms to run local applications, which intern reduces the cost and complexity at end points and as well as it deals effectively with heterogeneous devices when compared to manual operations. Suitable Gateway will be considered

depending on the nature of application. The following Fig. 3 shows the block diagram of basic Gateway.

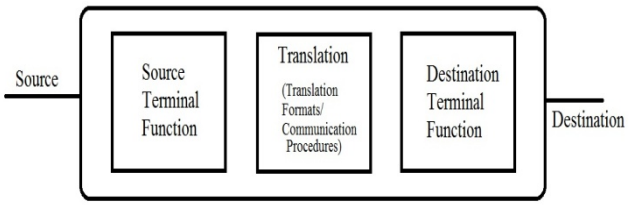


Fig. 3. Block diagram of Basic Gateway

Without Gateway noting happens in IoT. It connects the basic building blocks i.e. sensors which carries useful information to the IoT software. There are a lot of challenges in the implementation of Gateways for IoT [18]. Firstly it should be capable of maintaining huge amount of data producing from sensor networks, it should be able to handle data from variety of sensor networks, it is difficult to maintain permanent Gateways for IoT, privacy & security issues, the Gateway should have the capability to recognize the active sensors, Gateways should have coordination, and etc [19].

D. Cloud Services

The main purpose of the Cloud is to provide Internet based services. This service nature of Cloud makes it tightly coupled with IoT [20]. The Cloud has the capability to store, process and access the sensory data streams in IoT. Ultimately, the main objective of Cloud is to transfigure sensory data stream into productive information by applying intelligent decisions for the end user in cost-effective and optimal [21].

The basic architecture of the Cloud for IoT shown in the following Fig. 4. It has four basic phases Data Collection, Data Store, Data Analytics, and Application Processing, connectivity & Visualization.

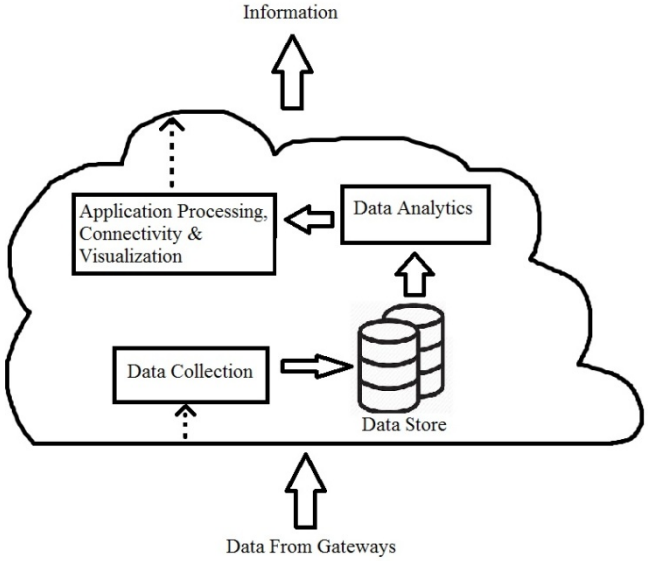


Fig. 4. Basic architecture of the Cloud for IoT

1) *Data Collection*: When Gateways forwards the data to Cloud, it collects the data in a systematic way. The collected data will be formatted according to the usage. In order to collect the data multiple times from the same source, it maintains data volume technique. When there are multiple sources, data from each source is separately collected [22]. The challenges in data collection of Cloud are In IoT, the Cloud should be able to maintain enough buffer space to collect the data as the data is having rapid growth and real world dynamic nature, it should be capable to maintain different types of data collected through different sources, it should be able to collect both structured and unstructured, it should be able to support numerous types of Gateways.

2) *Data Store:* The collected structured or unstructured data is stored dynamically and is categorized. The location of data that is stored in Cloud Storage area is forbidden but can be effectively accessed. It supports both SQL and NoSQL forms of data [23]. Data Storage of IoT leads to several challenges like identifying suitable database like SQL or NOSQL for the collected data according to application usage, maintaining of virtual directories for large volumes of data, space allocation issues for the massive volumes of data, and etc.,

3) *Data Analytics:* The revolution in modern Technologies leads to increase the availability of huge volumes of data, without applying proper analytics on the data makes in vain. So Data Analytics plays a crucial role in processing, examine, polish and model the data to successfully transform with the support of decision making by using Intelligent algorithms in IoT. We have availability of wide range of data analytical tools such as Weka, Excel, R, Hadoop, and etc [24]. Without Data Analysis, the services of IoT will not be fulfilling the needs of end user.

E. User Interactive Devices

At the end of IoT architecture Human – Smart Devices Interaction play a crucial role. The User Interactive Devices are the end points of IoT, which carries services to the user according to the requirements [25]. Based on the requests of the users, Smart devices interact with the cloud services and get well formatted information. Users provide requests through interfaces like touch screens, Gesture recognition tools, speech recognition tools, and etc. Inside the Cloud locale the requests are processed to generate appropriate services/data sent back to the client(s).

IV. BASIC ALGORITHM FOR IOT ARCHITECTURE

The following is the algorithm proposed for Basic IoT Architecture.

Algorithm: Basic IoT Architecture Algorithm

Input: *Data from different sensors: α, Filtered data at Gateways: β, and Cloud services: γ.*

Output: *n number of refined selected services in γ, m number of things.*

Procedure:

Start
Select *α data items from the sensors*
Sel_βi_data_items
Sel_Gateway(βi_data_items)
for i =1, 2, ... γ do
get *βi_data_items* set;
train M = filter
sel_i = selected features in *Mi* through *βi_data_items*
$n \leftarrow n + \{q, q \in \gamma\}$
m_get(n)
end

The proposed algorithm takes different type of data sets from a wide range of sensors. The Gateways select suitable data sets. These filtered data sets are forwarded to cloud services. Where after investigation, the model dynamically filters the data sets and prepares to provide the requested services among m things.

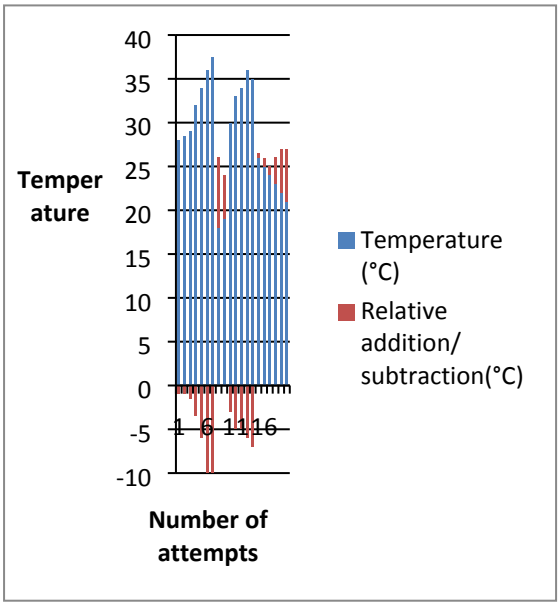
V. Working Nature of Proposed Algorithm on Real World Data

The proposed algorithm successfully implemented for a unit consists of sensors (Brainstrom Temperature Sensor), smart devices (Android Mobile) and a system (Intel Atom Processor S1200 Product Family for Microserver, Fedora Operating System). We supplied a wide range of temperatures through the sensor, at humid temperatures the system automatically adjusts room temperature by using Air Conditioning system and it indicates the message though smart device. The experiment results and the actions are mentioned in the below TABLE 1.

TABLE I
Temperature and Corresponding Actions

Experiment trail No.	Temperature(°C)	Automatic adjustment of temperature(°C)
1	28	-1
2	28.5	-1
3	29	-1.5
4	32	-3.5
5	34	-6
6	36	-10
7	37.5	-10
8	18	+8
9	19	+5
10	30	-3
11	33	-5
12	34	-5
13	36	-6
14	35	-7
15	26	+0.5
16	25	+1
17	24	+1
18	23	+3
19	22	+5
20	21	+6
	28	-1

The above TABLE 1 shows the adjustments of the current temperature as per the current comfortable body temperature by keeping humidity as a factor. For a given temperature, the system automatically adjusted the temperature to human comfortable zone. The below figure 5 shows the graphical representation of relative temperature changes. The red coloured region in the bar is the automatically adjusted relative value.



VI. Conclusion

The Internet of Things provides the best services that an average person would ever think. To implement these kind of services, the IoT system needs to be well defined in its architecture. The above proposed algorithm is best suitable to the basic architecture of IoT. The algorithm worked well for a simple kind of system and generated better results. This algorithm can be useful to implement future systems with numerous services.

References

[1] Xia, Feng, et al., "Internet of things," International Journal of Communication Systems 25.9 (2012): 1101.
[2] Doukas, Charalampos, and Ilias Maglogiannis, "Bringing IoT and cloud computing towards pervasive healthcare," Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on. IEEE, 2012.
[3] Caragliu, Andrea, Chiara Del Bo, and Peter Nijkamp, "Smart cities in Europe," Journal of urban technology 18.2 (2011): 65-82.

[4] Silviu C. Folea and George Mois, "A Low-Power Wireless Sensor for Online Ambient Monitoring," *Sensors Journal IEEE*, vol. 15, pp. 742-749, 2015, ISSN 1530-437X.
[5] Miorandi, Daniele, et al., "Internet of things: Vision, applications and research challenges," Ad Hoc Networks 10.7 (2012): 1497-1516.
[6] Wyburn, George M., Ralph W. Pickford, and Rodney Julian Hirst, "Human senses and perception," (1964).
[7] Sugawara, Yoshiaki, et al., "Use of human senses as sensors," Sensors 9.5 (2009): 3184-3204.
[8] Su, Xiang, et al., "Connecting IoT Sensors to Knowledge-Based Systems by Transforming SenML to RDF," Procedia Computer Science 32 (2014): 215-222.
[9] Gubbi, Jayavardhana, et al., "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Generation Computer Systems 29.7 (2013): 1645-1660.
[10] Kelly, Sean Dieter Tebje, Nagender Kumar Suryadevara, and Subhas Chandra Mukhopadhyay, "Towards the implementation of IoT for environmental condition monitoring in homes," IEEE Sensors Journal 13.10 (2013): 3846-3853.
[11] Qian, Zhihong, and Yijun Wang, "IoT technology and application," Acta Electronica Sinica 40.5 (2012): 1023-1028.
[12] Chi, Qingping, et al., "A reconfigurable smart sensor interface for industrial WSN in IoT environment," IEEE Transactions on Industrial Informatics 10.2 (2014): 1417-1425.
[13] Benbasat, Ari Y., Stacy J. Morris, and Joseph A. Paradiso, "A wireless modular sensor architecture and its application in on-shoe gait analysis," Sensors, 2003. Proceedings of IEEE. Vol. 2. IEEE, 2003.
[14] Estrin, Deborah, et al, "Next century challenges: Scalable coordination in sensor networks," Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking. ACM, 1999.
[15] Zhu, Qian, et al., "Iot gateway: Bridgingwireless sensor networks into internet of things," Embedded and Ubiquitous Computing (EUC), 2010 IEEE/IFIP 8th International Conference on. IEEE, 2010.
[16] Datta, Soumya Kanti, Christian Bonnet, and Navid Nikaein, "An IoT gateway centric architecture to provide novel M2M services," Internet of Things (WF-IoT), 2014 IEEE World Forum on. IEEE, 2014.
[17] Rahmani, Amir-Mohammad, et al., "Smart e-health gateway: Bringing intelligence to internet-of-things based ubiquitous healthcare systems," 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC). IEEE, 2015.
[18] Mainetti, Luca, Luigi Patrono, and Antonio Vilei, "Evolution of wireless sensor networks towards the internet of things: A survey," Software, Telecommunications and Computer Networks (SoftCOM), 2011 19th International Conference on. IEEE, 2011.
[19] Sheng, Zhengguo, et al., "A survey on the ietf protocol suite for the internet of things: Standards, challenges, and opportunities," IEEE Wireless Communications 20.6 (2013): 91-98.
[20] He, Wu, Gongjun Yan, and Li Da Xu, "Developing vehicular data cloud services in the IoT environment," IEEE Transactions on Industrial Informatics 10.2 (2014): 1587-1595.
[21] Calheiros, Rodrigo N., and Rajkumar Buyya, "Cost-effective provisioning and scheduling of deadline-constrained applications in hybrid clouds,"International Conference on Web Information Systems Engineering. Springer Berlin Heidelberg, 2012.
[22] Rolim, Carlos Oberdan, et al., "A cloud computing solution for patient's data collection in health care institutions," eHealth, Telemedicine, and Social Medicine, 2010. ETELEMED'10. Second International Conference on. IEEE, 2010.
[23] Wang, Cong, et al., "Privacy-preserving public auditing for data storage security in cloud computing," INFOCOM, 2010 Proceedings IEEE. Ieee, 2010.
[24] Talia, Domenico, "Toward cloud-based big-data analytics," IEEE Computer Science (2013): 98-101.
[25] Carroll, David W., et al., "Interactive devices and methods," U.S. Patent No. 6,285,757. 4 Sep. 2001.

Dynamic Search for Spatio-Textual Queries on Location Based Applications

K.Haritha[1], M.Vani Pujitha [2*]

[1]*PG Scholar, Dept. of CSE, V.R Siddhartha Engineering College,* [1]E-mail:kilaruharitha@gmail.com
[2*] *Assistant Professor, Dept. of CSE, V.R Siddhartha Engineering College,* [2*]E-mail:pujitha.vani@gmail.com

***Abstract*—In an entity search over spatial data, a user specifies requirements in the form of a query, and the main task is to find a route to a target object that goes via geographical locations while satisfying the search specifications. For example, consider a tourist in a foreign city wants to find a restaurant from their current location by any means of transportation. Elaborating the search with further filters such as restaurant type (veg, non veg), menu items specifications etc would yield much better results. Although prior approaches formulated a way to integrate these filters into specific target object search, they may tend not to be useful to the user with respect to dynamic perspectives since the system is predefined with static filters. In realistic scenarios, the navigational service provider should consider additional complicating factors such as the working hours of the entities to be visited, type of service those entities cater to and the possible restrictions on the order by which those entities may be visited, possible change of items they service for .These factors are called as temporal constraints. Incorporation of such temporal constraints in spatial scenario leads to a new spatial temporal approach to target object queries. Temporal approximation algorithm is used for target object search over spatial data to handle temporal constraints on them.**

***Keywords*- Spatial, Temporal, Constraints**

I . INTRODUCTION

Geographical information system (GIS) or Geospatial information system, is any system for capturing, storing, analyzing, managing and presenting data and associated attributes which are spatially referenced to earth

A Geographic information system (GIS) is not one particular component, nor a single analysis but rather a collection of hardware, software, data organisations, and professionals that together help people to represent and analyse geographic data

These days mobile devices are using geographical web. A fundamental service in geo web is spatio textual query that takes user location and keyword set as inputs and returns the most spatially and textually relevant objects.

The objective of the project is to find a optimal route query which considers all the necessary route constraints defined by the user. The constraints may be arbitrary constraints such as an ATM machine must be visited before a restaurant or temporal constraints such as working hours of a restaurant.

The project has three modules. In the first module a source location and a destination location are given. Shortest path between source and destination are found using shortest path algorithms. In the second module, a source location is given, and the other locations which must be visited before reaching target location are also given. Shortest path between source and these intermediate locations is calculated. This shortest path is the result. For example if a user is in location A and before reaching target location B user should go to restaurant. So the system should find out the shortest path between source and the restaurant and then give result accordingly. This is not applicable when the user wants to add additional features in searching. For example the working hours of the restaurant and the menu items of the restaurant are changing from time to time. These additional features are called as temporal constraints.

In the third module the additional features required by the user are satisfied.

Geocoding is a feature of all geospatial applications. Geocoding converts a street address to a latitude or longitude position so it can be accurately placed in a map. The opposite function of the geocoding is reverse geocoding. It is the process of deriving the location of the nearest road segment to a point with specified latitude or longitude. The derived information which includes world coordinates, address location and directional distances from reference points can then be used for routing, searching of points of interest.

Main advantage of this technique is when a person visits an unknown place and wants to find nearest restaurant that serves biryani. Here the user specified both spatial and temporal constraints. Suppose the menu of the restaurant changes day to day. So the system should consider these temporal constraints as well.

The other advantage is any number of locations can be taken dynamically. There is a dynamic indexing tree structure called as KcR tree (Keyword count R tree) which grows dynamically. KcR tree consists of root node, leaf nodes and middle nodes. KcR tree consists of various locations information in the form of x coordinate and y coordinate and the distances from one location to the other location. It is a dynamically growing tree structure. It consists of root node, areas node, POI node, spatial links node.

Achievements made in this paper are listed as below:

A KcR indexing trees structure is developed. With this searching is made easy

The locations information is changing time to time. So the information is updated in a minute wise manner. This updating is represented in the form of a table. The structure of a KcR tree is shown below:

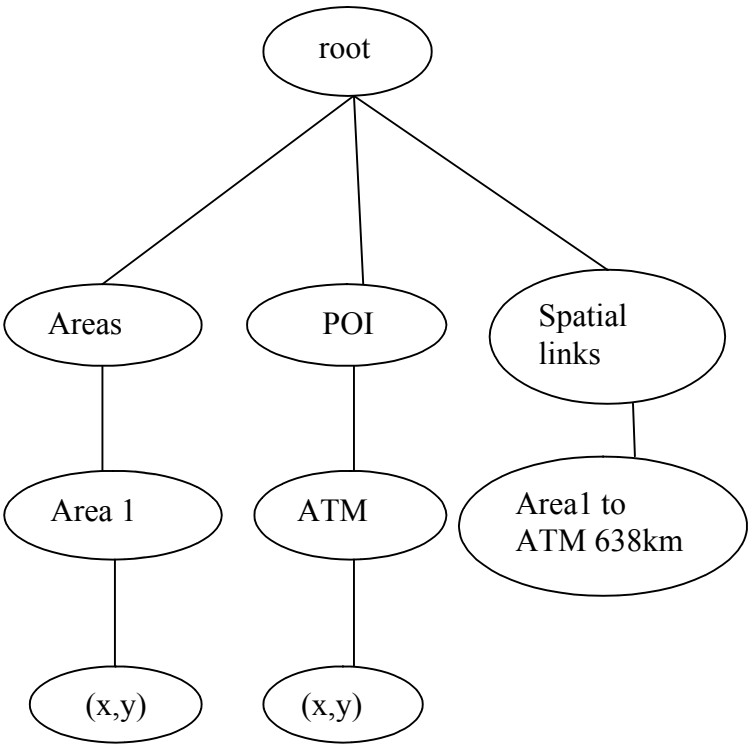


Figure 1: KcR tree

II. RELATED WORK

In the existing methods Xin Cao, Gao Cong, Christian S.Jensen[1] proposed a Location-aware top-k Prestige-based Text retrieval (LkPT) query, that retrieves the top-k spatial objects ranked according to prestige based similarity and nearest location. Here two algorithms Early Stop EBC Algorithm and Sub graph-Based EBC Algorithm are used. The advantage of this method is though the object does not contain the query terms the object can be identified as relevant. The disadvantage of this method is computation of LkPT query is expensive due to PR scores

Goerge Tsatsanifos and Akrivi Vlachou[2] proposed a technique called top-k spatio textual preference query that retrieves a set of objects ranked according to the richness of the maintenance in their neighbouring objects. Here Spatio Textual Preference Search algorithm is used. To further improve the performance an indexing technique called SRT-index is developed. In STPS algorithm highly ranked objects are retrieved first and then its neighbouring objects are searched. The main disadvantage of this method is determining efficiently the best feature objects from all feature sets that do not violate the spatial constraint.

Mingdong Zhu, Derong Shen, Ling Liu and Ge yu[3] proposed a method called Locality Sensitive Hashing. LSH is a solution for k Nearest Neighbours' problem. Here multi dimensional objects are hashed so that similar objects will get the same hash value. The main disadvantage of this method is that LSH works well for sparse area (the area which has less number of nearest neighbour objects) and it works hard on area with more number of nearest neighbour objects. It leads to computational bottleneck.

Muhammad Aamir Cheema, Wenjie Zhang, Xuemin Lin, Ying Zhang[4] proposed a technique. In this technique influence zones are calculated for each and every query points by using algorithms. A set of objects are given and a query t is given, then a point o is called the RkNN of t if t is one of the k closest objects of o. Main advantages of influence zone includes location based applications, market applications and decision making systems.

Hideki Sato Ryoichi Narita [5] proposed the concept of Regular Polygon based Search Algorithm (RPSA). The main advantage of this technique is when a mobile user wants to access location information from management server. If the databases are non local then the user cannot get the location information.

Lia Chen, Jianliang Xu, Xin Lin, Christian S.Jensen, Haibo Hu[6] proposed Bound and Prune algorithm. The advantage of this method is missing objects will appear in result. The drawback of this method is only the object keywords are considered.

Christodoulos Efstathiades, Alexandros Belesiotis, Dimitrios Skoutas proposed solution to Spatio-Textual Point-Set Join query problem. Now days, people are posting in online networking sites. Main advantage STPS Join is it finds user who are exhibiting same "geo-textual" behaviour. The disadvantage with this method is it requires verification phase so the execution time is more for this method

R.Subbarao and K.Sri kanth proposed a cache based approach. The main advantage of this method is it returns required number of accurate results .The disadvantage of this method is it is more costly

III. PROPOSED SYSTEM

The main advantage of proposed system is user can add additional features while searching the route from source to destination. These additional features can be temporal constraints like the working hours of an ATM, various menu items in a restaurant.

The proposed system is implemented in the following steps:

Input: start location a target location b search queries T_1, T_2,.....,T_m ordered according to C, a dataset d an order α over D

Output: shortest path from source to the destination that satisfies the user constraints

Method: Optimal Approximation Algorithm

Step 1: Place repository set up

Step 2: Route repository set up

Step 3: Querying user interface
Step 4: Conversion to comma separated values

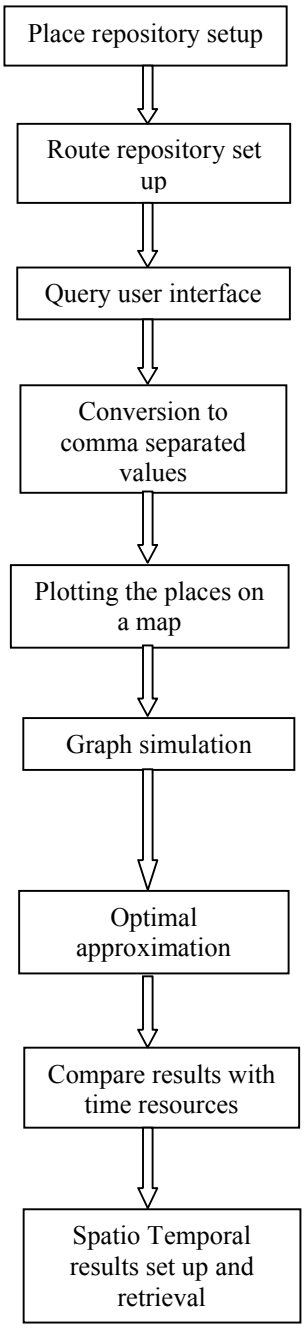
Step 5: Plotting the places on a map

Step 6: Graph simulation

Step 7: Optimal Approximation Algorithm

Step 8: Compare results with time resources

Step 9:Spatio Temporal results set up and retrieval



A. Places Repository set up

In this phase location information is accumulated with respect to latitude(x-coordinate) and longitude(y-coordinate) in the form of sql queries. An example query is shown below
insert into route1(route name, x coordinate, y coordinate) values ('area1','123','456');

B. Route Repository Setup

In this phase the distance between two locations is specified in the form of sql queries. An example query is given below
insert into route1map(place1, place2, distance) values('area1', 'area2', '319');

C. Query User Interface

By using this interface the user can select source and destination along with the required filters search. Best path is obtained between source and destination. The interface is developed using java swings. The interface contains textboxes, checkboxes, submit buttons and reset buttons. The interface contains corresponding map with locations

D. Conversion into comma separated values

In this phase database values are converted into comma separated values. These csv values are used in optimal path mining queries

E. Plotting the places on a map

In this phase the locations are plotted on a given map interface along with paths in the form of weights or distances between them.

F. Graph simulation

Here all routes from source to destination are displayed. The best route from source destination is also given. The time taken to display the result is also given. Result is displayed in browser.

G. Optimal approximation algorithm implementation

Here the search is based on temporal constraints as well. By considering temporal constraints the path from source to destination is identified.

The shortest path between src and dest are found as follows
Let a_1, a_2.....,a_n be the locations and r_1,r_2,.....r_n be the points of interest which contains the user defined constraints
Now the shortest distance between two points is found as follows compute the shortest distance between each and every nodes by using Floyd war shall algorithm

1For i=1 to n

2. For j=1 to n

3. Initialize dist(a[i],a[j])=infinity

4. For i=1 to n

5. Initialize dist(a[i], a[i])=0

6. for each edge a[i],a[j]

7. Dist (a[i],a[j])=weight(a[i],a[j])

8. for k from 1 to p

9. for m from 1 to n

10. for c from 1 to m

11. If dist (a[m],a[c])> 12.dist(a[m],a[k])+dist(a[k],a[c])

13. Then

14.dist(a[m],a[c]) = dist(a[m],a[k])+dist(a[k],a[c])

15. end if

16. Similarly find the shortest distance between $r_1, r_2, \ldots, r_n$ points as well and finally obtain the path from source to destination points

F. Comparing results with time resources

Here the results are compared based on time

G. Results set up and retrieval

Here the output of is displayed to the user

IV. EXPERIMENTAL ANALYSIS

A. *Dataset*

Here the dataset is location dataset in a particular city. Location dataset contains various places in that particular city and the distances from one place to another place. The data set is in the form of csv values

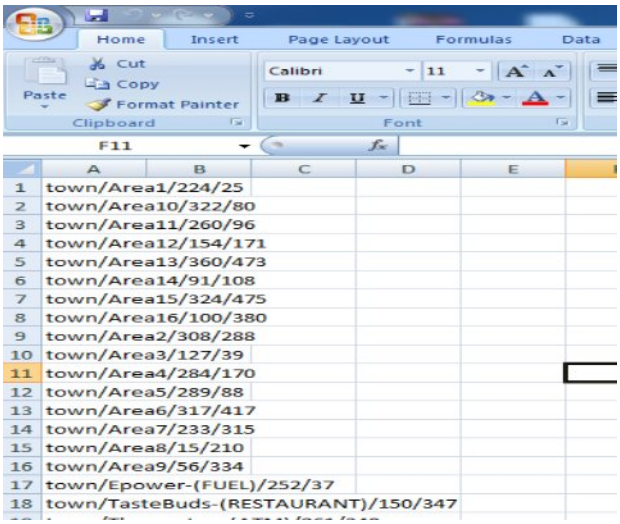


Figure 2: Dataset

The locations are represented on a map. There are total 11 maps. User can select according to their wish

Figure 3:Maps

When the user selects existing system the following map is obtained

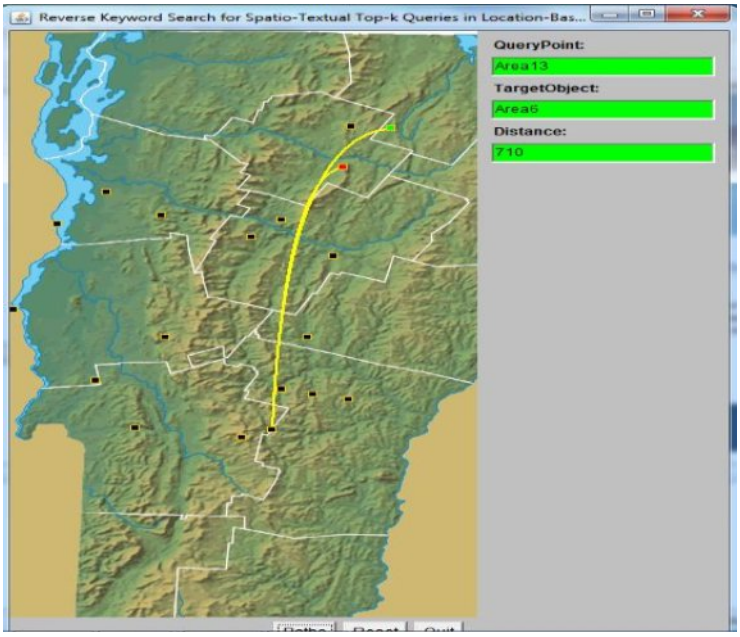


Figure 4: Output Map between source and target

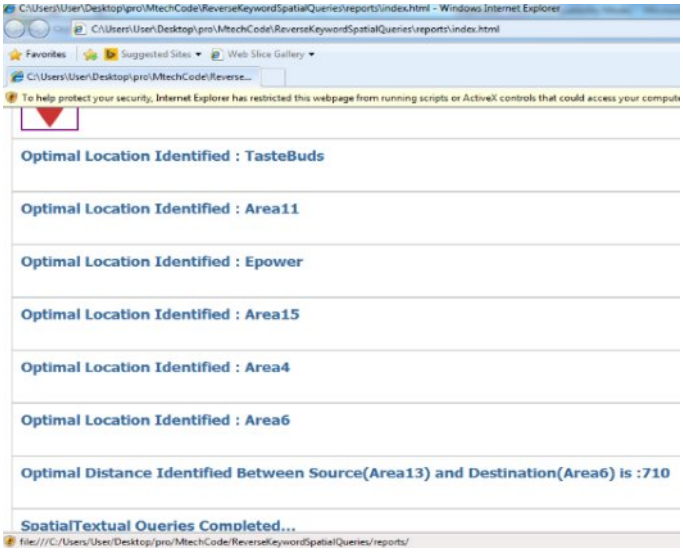


Figure 5: Distances between source and target

When the user selects proposed system then the following visual KcR tree indexing is obtained. It contains locations information in the form of x coordinate and y coordinate. It is a tree developed to handle dynamic growth

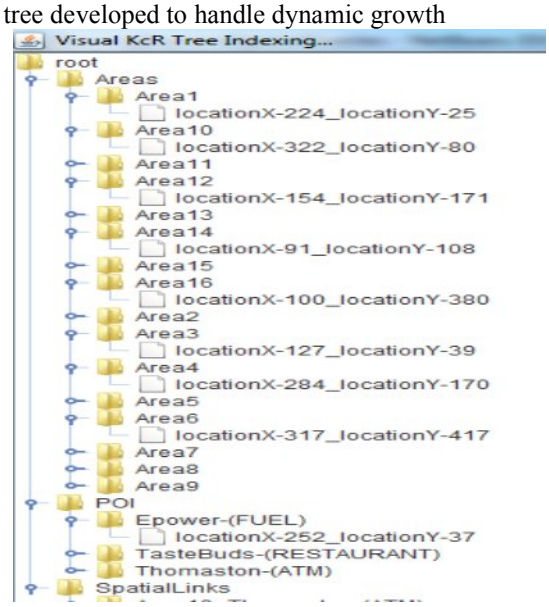


Figure 6: KcR tree

In the proposed system the user enters the constraints like restuarant, ATM, fuel station

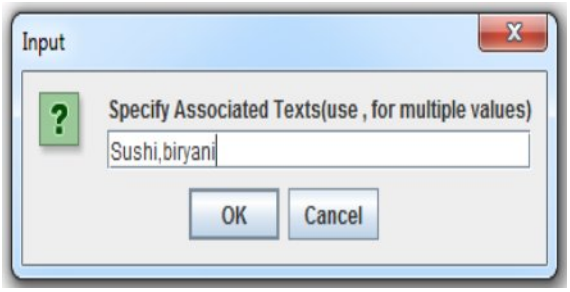


Figure 7: User enters constraints

Figure 8: output for choice found

When the user selects enhancement then the information which is changing time to time is represented in the following table

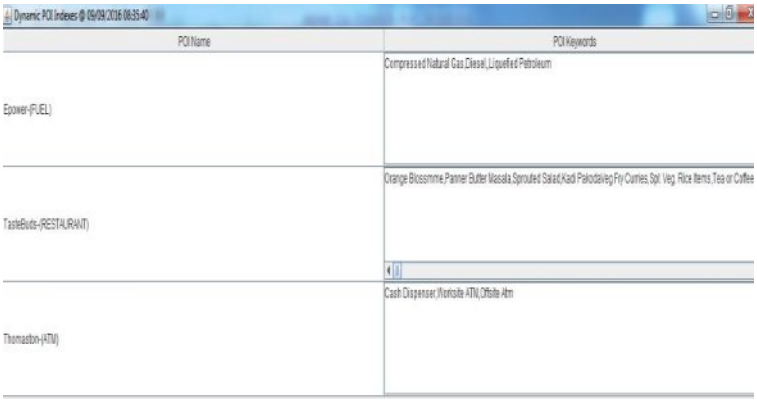


Figure 9: Temporal constraints

V . CONCLUSION

The proposed algorithm gives information about the user defined constraints and optimal location effectively. A combination of two algorithms reverse Forward Search and temporal approximation algorithm is used which efficiently implements the optimal query system. The constraints can be arbitrary or temporal. In future an optimal query system which can handle more categories to be visited should be developed. And if there is no category or place found in the selected route it displays no places exist in the required route. So there is a need to develop the system which can give alternative best path between source and destination.

REFERENCES

[1] X.Cao, G.Cong and C.S.Jensen, " Retrieving top-k prestige based Relevant spatial web objects" VLDB 3(1): 373-384, 2010

[2] George Tsatsanifos and Akrivi Vlachou, "On Processing Top-k Spatio-Textual Preference Queries", Open Proceedings, Mar.2015

[3] Mingdong Zhu, Derong Shen, Ling Liu and Ge Yu, "Hybrid-LSH for Spatio-Textual Similarity Queries" 2014

[4] Muhammad Aamir Cheema, Wenjie Zhang, Xuemin Lin, Ying Zhang, "Efficiently processing snapshot and continuous reverse k nearest Neighbours queries", VLDB, 2012

[5] Hideki Sato and Ryoichi Narita, "Efficient Maximum Range Search on Remote Spatial Databases Using k-Nearest Neighbor Queries", ScienceDirect, 2013

[6] Lia Chen, Jianliang Xu, Xin Lin, Christian S.Jensen, Haibo Hu, "Answering Why-Not Questions on Spatial Keyword Top-k Queries"

[7] Christodoulos Efstathiades, Alexandros Belesiotis, Dimitrios Skoutas, "Similarity Search on Spatio-Textual Point Sets", open proceedings,2016
.

[8] R.Subbarao and K.Sri kanth, "A Novel Geo-coding and Cache Based Approaches for Spatial Queries", IJETT, Oct.2013

[9] Xiaoling Zhou, Wei Wang, Jianliang Xu, "General Purpose Index-Based Method for Efficient MaxRS Query", DEXA (1), 2016

[10] Dingming Wu, Byron Choi, Jianliang Xu, Christian S.Jensen, "Authentication of moving Top-k Spatial Keyword Queries",IEEE Trans. Know, 2015.

HEART DISEASE DIAGNOSIS USING PREDICTIVE DATA MINING

M.Swathi Lakshmi
MTech Student,
SRKIT,Vijayawada,India
marisettyswathi@gmail.com

Dr D.Haritha
Professor in CSE Department
SRKIT,Vijayawada,India
harithadasari@rediffmail.com

Abstract

With the fast increasing rates of heart disease or Cardiovascular Disease the classification and prediction of heart diseases is of significant work. Computerized classification of heart diseases is more useful for the physicians for fast diagnosis. Prediction of heart disease accurately can help in saving the patients lives. In this paper various existing data mining techniques are applied, analyzed for classification and prediction of heart disease. The dataset used is the Cleveland Heart Database taken from UCI learning data set repository. The findings of this study revealed all the models built from Naïve Bayes classifier and SVM have high classification accuracy and are generally comparable in predicting heart disease cases.

Keywords: ANN, Data Mining, ROC.

I.INTRODUCTION:

Health is rooted in everyday life. The healthcare industry is one of the world's largest and fastest-growing industries and is the back bone of our society. For this industry a major challenge is providing quality services at reasonable prices. Diagnosing a patient's condition accurately with appropriate treatment, monitoring and evaluating the effectiveness of the treatment on a regular basis is one of such quality service. At present, the number of people suffering with heart disease is on a rise. **Cardiovascular disease** (**CVD**) is a class of the heart or blood vessel diseases. Cardiovascular disease includes coronary artery diseases (CAD) like heart attack, stroke, angina, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, atrial fibrillation, congenital heart disease,8].This leads to number of deaths these days. At an early stage proper diagnosis is very crucial task. The advancement of information technology, software development and system integration facilitated the development of multifaceted computer systems. One of such new powerful technology is Data mining. Data mining, deals with the extraction of hidden information from large databases, with great potential to help companies to focus on the most important information present inherently in their data warehouses. Data mining tools allow organizations to make proactive, knowledge-driven decisions based on the prediction of future trends and behaviors. The automated, prospective analyses offered by data mining outraged the analyses provided by retrospective tools typical of decision support systems. Data mining tools can answer several questions that traditionally were too time consuming to resolve in less time in an efficient manner.

Section II briefs the previous work done in the area of Heart Disease Prediction using Data mining techniques. Section III describes various Data mining techniques. The architecture of using Data mining techniques for heart disease prediction is explained in Section IV. The results and efficiency of the various techniques are discussed in Section V.

II.LITERATURE SURVEY

In 2015 T.Georgeena et al., applied Apriori Algorithm Heart Disease Diagnosis System Using Apriori Algorithm". Where k-item sets are used to explore (k+1) item sets. The data will judge the efficiency and correction rate of the algorithm. They have reduced six attributes to four which will be employed for the prediction of heart conditions [14].

In 2014 B.Venkatalakshmi et al. applied Predictive data mining techniques Naïve Bayes and Decision

Tree for the prediction of heart disease on WEKA tool. The author concluded that Naïve Bayes method outperforms Decision Tree in prediction [11].

In [12] Deepali Chandna showed how information gain method, feature selection technique with k-nearest neighbor's algorithm can be used along with adaptive neuro fuzzy inference systems in diagnosing new patient cases. The study found that the accuracy for the proposed approach is 98.24% for diagnosing patients for heart diseases.

In 2013 Deepti Vadicherla et al. applied the two algorithms namely Support Vector machine and Artificial Neural Network (ANN) to diagnose the heart disease, and shown that results of the SVM classification algorithm compared to the ANN classification are encouraging [8].
In 2011 K. Usha Rani applied neural networks for prediction and analysis of heart disease [13].Backpropogation algorithm with variable learning rate is used to train the networks and shown the Classification efficiency as 90% approximately.

III. RELATED WORK

In this section all the data mining techniques that are used in our analysis are briefly discussed. The architecture for prediction of heart disease model is shown in Figure1.

A. Naïve Bayes:

The Naïve Bayes algorithm is used for classification based on Bayes rule, that assumes all the attributes $X_1,\ldots,X_n$ are conditionally and mutually independent given Y. The value of this assumption dramatically simplifies and reduces the complexity and representation of $P(X|Y)$ and the problem of estimating it from the training data. Considering the case where $X = (X_1, X_2)$[1].

$$P(X|Y) = P(X1, X2|Y) = P(X1|X2, Y)P(X2|Y) = P(X1|Y)P(X2|Y)$$

This can be represented as

$$P(X1\ldots..Xn|Y) = \prod$$

B. Support Vector Machine:

The Support Vector Machine (SVM) proposed by Vapnik has greater potential in the machine learning research. Several recent studies have reported that the SVM (support vector machines) are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVMs are set of related supervised learning methods used for classification and regression. SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed [8]. Two parallel hyperplanes are realized on each side of the hyperplane that separates the data such that the distance between the two parallel hyperplanes is maximized. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be.[8].

C. Neural Net:

A neural network is a system of programs and data structures that approximates the operation of the human brain. A neural network comprises a large number of processors in parallel, each with its own local knowledge and access to data in its local memory. Typically, a neural network is initially "trained" or fed with large volume of data and data relationship rules (for example, "A grandfather is older than a person's father") [13].

D .Rule Induction:

The rules induced from examples are represented as logical expressions of the following form:[5]

IF (*conditions*) *THEN* (*decision class*)*;*

where conditions are conjunctions of elementary tests on values of attributes, and decision indicates the assignment of an object (which satisfies the condition part) to a given decision class. This operator learns a pruned set of rules with respect to the information gain from the given Example Set.

E. Random Forest: Random Forest developed by Leo Breiman is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. By aggregating i.e. taking the majority vote for classification or average for regression from the predictions of the ensemble the predication is made [9].

F.K-Nearest Neighbor: The k-Nearest Neighbor algorithm is based on learning by analogy, that is, by comparing the test example with the similar training examples. The training examples are described by n attributes. In an n-dimensional space each example

represents a point. In this way, all of the training examples are stored in an n-dimensional pattern space. A k-nearest neighbor algorithm searches the pattern space for the k training closer examples to the given unknown example. These k training examples are the k "nearest neighbors" of the unknown example. "Closeness" is found with the Euclidean distance [4].

G. Decision Trees: Decision Trees (DTs) is a supervised learning method used for classification. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred [2]. The benefits of decision tree in data mining are ability to handle variety of input data such as nominal, numeric and textual , ability to process the dataset that contain the errors and missing values and available in various packages of data mining and number of platform.

H. Decision Stump: The Decision Stump operator is used for generating a decision tree with only one single split that can be used for classifying new examples. This operator can be very efficient when boosted with operators like the AdaBoost operator [3].

F. Random Tree: The Random Tree operator works exactly like the Decision Tree operator with one exception i.e. for each split only a random subset of attributes is available [7].

G.CHAID: The CHAID decision tree operator works similarly as the Decision Tree operator with one exception that it uses a chi-squared based criterion instead of the gain ratio. Moreover this operator cannot be applied on Sets with numerical attributes [6].

H. Genetic algorithm: Evolutionary computing started by lifting ideas from biological theory into computer science. Genetic algorithms are predominantly used in evolutionary computing. Evolutionary algorithms are used in problems for optimization. These require a data structure to represent and evaluate solution from old solutions [10].

I.**ROC curve** : It is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is the true positive rate (TPR) plot against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as (1 - specificity).

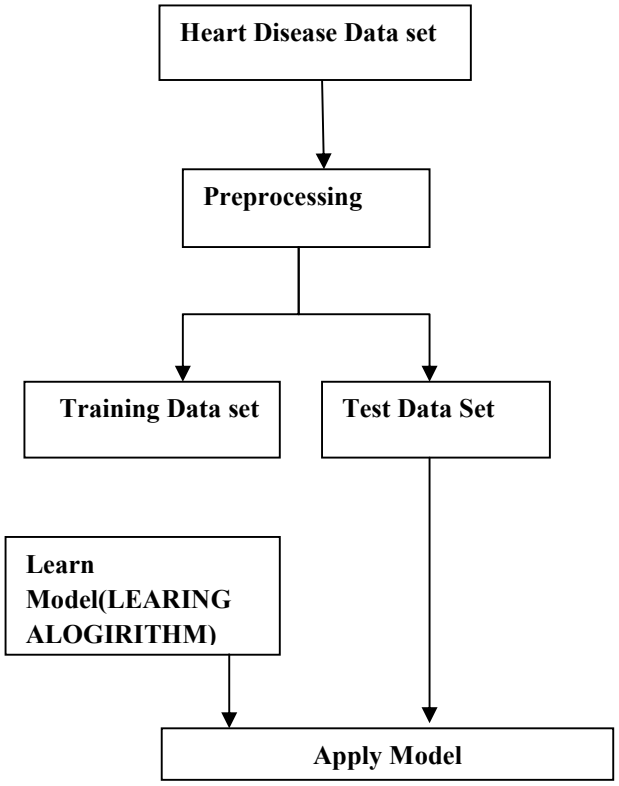


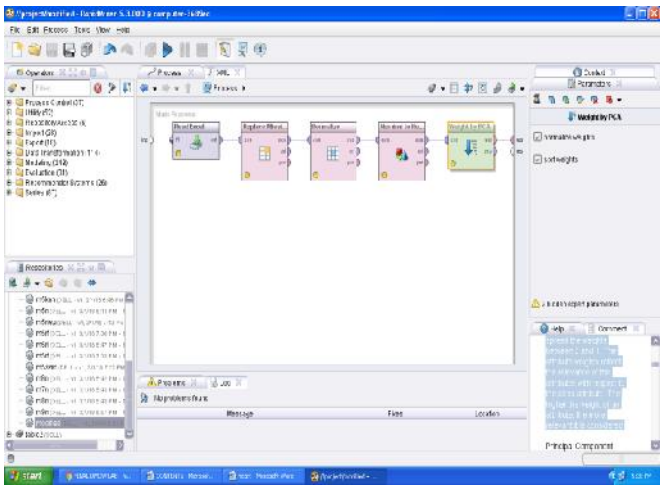
Fig 1: Architecture of Heart Disease Prediction Model

IV Results and Discussions

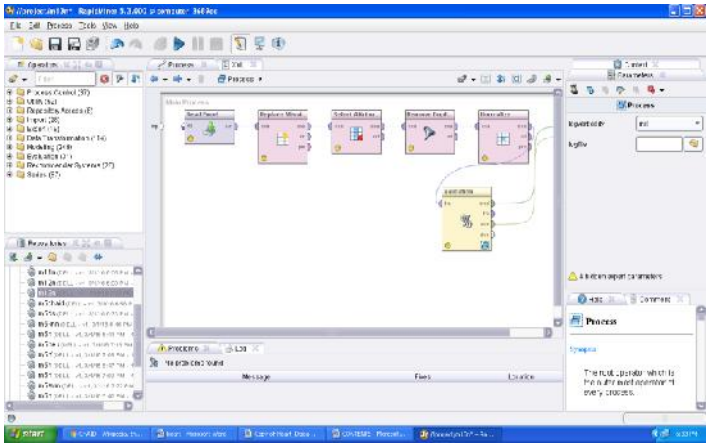
In this section we describe the dataset taken for experimentation for finding the accuracy comparison of all the algorithms described above in heart disease analysis and prediction.

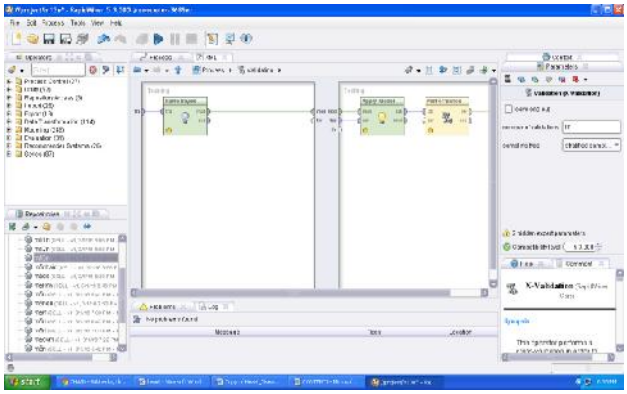
The Data set can be downloaded from this UCI machine repository and data set is created by Hungarian Institute of Cardiology. The attributes that are useful in prediction are described in (Table1).This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them and the characteristics of dataset is multi variant and attribute types are categorical, Integer, Real.

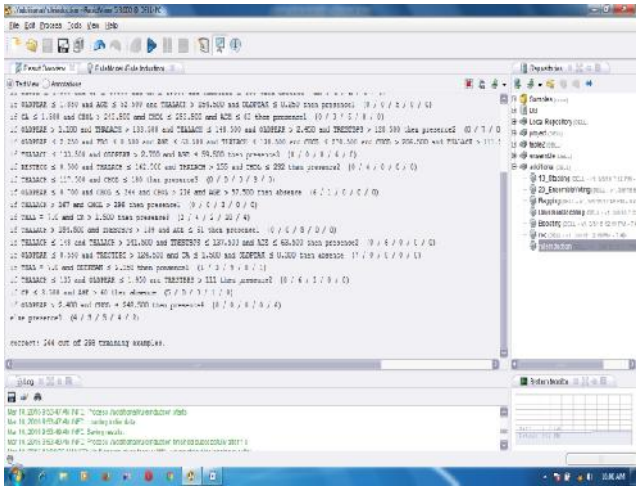
The Figure 2 depicts the process of applying PCA for feature Selection in Rapid Miner . Initially, the input data set is taken, and then it is converted into numerical type data using a component called Nominal to Numeric. Before applying the resulting



Attribute Index	Selected attributes
3	age: age in years
4	sex: sex (1 = male; 0 = female)
9	cp:chest pain type --Value 1:typical angina --Value 2:atypical angina --Value 3:non-anginal pain -- Value 4: asymptomatic
10	Trestbps: resting blood pressure (in mm Hg on admission to the hospital)
12	chol: serum cholestoral in mg/dl
16	fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
19	restecg: resting electrocardiographic results --Value 0:normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
32	thalach: maximum heart rate achieved
38	exang: exercise induced angina (1 = yes; 0 = no)
40	oldpeak = ST depression induced by exercise relative to rest
41	slope: the slope of the peak exercise ST segment
44	Ca: number of major vessels (0-3) colored by flourosopy
51	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
58	Num(The predicted attribute)







ALGORITHM	ACCURACY
Naïve Bayes	84.87
Decision Tree	72.61
Decision stump	69.68
K-NN	75.96
Rule Induction	76.19
CHAID	54.14
Random Tree	65.35
Random Forest	75.97
Neural Net	78.24
SVM	84.19

V Conclusion

Prediction of heart disease accurately can help in saving the patients lives. In this paper we have used various data mining techniques to build a model which is capable of accurately predicting the existence of the heart disease. From the result we can say that Naïve Bayes and SVM perform well in prediction and detection of the heart disease. This work can be extended further by using ensemble classifiers to build the model which recognizes the heart disease even more accurately.

References:

[1]. Tina R. Patil, Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification" International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013.

[2]. Shahrukh Teli, Prashasti Kanikar, "A Survey on Decision Tree Based Approaches in Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.

[3]. Ayinde A.Q, Dr Adetunji A.B, Bello M and Odeniyi O.A, "Performance Evaluation of Naive Bayes and Decision Stump Algorithms in Mining Students' Educational Data", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No 1, July 2013.

[4]. M.Akhil jabbar B.L Deekshatulua Priti Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.

[5]. Jerzy Stefanowski and Sławomir Nowaczyk, "An Experimental Study of Using Rule Induction Algorithm in Combiner Multiple Classifier", International Journal of Computational Intelligence Research, Vol.2, No.X (2006).

[6]. M. Ramaswami and R. Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010.

[7].http://docs.rapidminer.com/studio/operators/modeling/predictive/trees/random_tree.html.

[8].DeeptiVadicherla, Sheetal Sonawane "Classification Of Heart Disease Using Svm And ANN", International Journal of Research in Computer and Communication Technology, Vol 2, Issue 9, September -2013.

[9]. JehadAli1 ,Rehanullah Khan2 , Nasir Ahmad3 , Imran Maqsood "Random Forests and Decision Trees" , IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[10]. Gentic Algorithms in search, Optimization and Machine Learing 1st Edition, By David E.Goldberg,, Pearson Edition..

[11].B.Venkatalakshmi, M.V Shivsankar, "Heart Disease Diagnosis Using Predictive Data Mining", International Journal of Innovative Research in Science, Engineering and Technology, vol.3, pp. 1873–1877, Mar. 2014

[12] Deepali Chandna, "Diagnosis of Heart Disease Using Data Mining Algorithm", International Journal of Computer Science and Information Technologies, Vol. 5 (2) , PP 1678-1680,2014.

[13]K. Usha Rani, "Analysis of Heart Diseases Dataset using Neural Network Approach", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011.

[14] T.Georgeena.S. Thomas, Siddhesh.S. Budhkar, Siddhesh.K. Cheulkar, Akshay.B.Choudhary, Rohan Singh," Heart Disease Diagnosis System Using Apriori Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 2, February 2015

Healthcare applications of the Internet of Things (IoT): A Review

M.V.D.N.S.Madhavi[a], K.Hemalatha[a], P.V.S.Sairam[b] and D.Rajani[a]

[a]Department of Mathematics, V.R.Siddhartha Engineering College, Vijayawada

[b]Department of Physics, Andhra Loyola College, Vijayawada

Abstract

The fields of computer science and electronics are combined to result one of the most prominent technological advances in the form of realization of the Internet of Things (IoT). Among many of applications enabled by the Internet of Things (IoT), smart and connected health care is a particularly important one. The impact of IoT in healthcare, although still in its early stages of development has been significant. Networked sensors, either worn on the body or embedded in our living environments, make possible the gathering of rich information indicative of our physical and mental health. This paper tries to review and comprehend the applications of IoT in custom-made healthcare to achieve excellent healthcare at affordable costs. We have explained in brief what IoT is, how it functions and how it is used in combination with wireless and sensing techniques to implement the desired healthcare applications. Here, we highlighted the opportunities and challenges for IoT in realizing the vision of the future of health care.

Keywords*— Internet of things, healthcare, remote health monitoring, visualization, computing, wireless sensor networks etc.*

Introduction

The term Internet talk about to wide category of applications and protocols built on top of refined and interconnected computer networks, serving trillions of users around the world round the clock. Infact, we are at the beginning of an emerging era where global communication and connectivity is neither a dream nor a challenge anymore. Consequently, the focus has shifted toward a one-piece integration of people and devices to converge the physical realm with human made virtual environment, creating the so called Internet of Things. This phenomena reveals two important pillars of IoT i.e., Internet and Things, which require more clarification. Every object capable of connecting to the Internet will fall into the "Things" category which includes the generic entities like smart devices, sensors, human beings and other objects which able to communicate with other entities and accessible anytime, anywhere

In 2008 the number of things connected to the Internet was greater than the people living on Earth.Within 2020 the number of things connected to the Internet will be about 50 billion. In the IoT, 'things' are expected to become active participants in business, information and social processes where they are enabled to interact and communicate among themselves and with the environment by exchanging data and information 'sensed' about the environment,

while reacting autonomously to the 'real/physical world' events and influencing it by running processes that trigger actions and create services with or without direct human intervention. Interfaces in the form of services facilitate interactions with these 'smart things' over the Internet, query and change their state and any information associated with them, taking into account security and privacy issues. The internet of things (IoT) is the internetworking of physical devices, vehicles, buildings and other items—embedded with electronics, software, sensors, actuators, and network connectivity that enable these objects to collect and exchange data. In 2013 the Global Standards Initiative on Internet of Things (IoT-GSI) defined the IoT as "the infrastructure of the information society".The IoT allows objects to be sensed and/or controlled remotely across existing network infrastructure, creating opportunities for more direct integration of the physical world into computer-based systems, and resulting in improved efficiency, accuracy and economic benefit. The availability of data till now at unimagined scales and temporal longitudes coupled with a new generation of intelligent processing algorithms can: (a) facilitate an evolution in the practice of medicine, from the current post facto diagnose-and treat reactive paradigm, to a proactive framework for prognosis of diseases at an incipient stage, coupled with prevention, cure, and overall management of health instead of disease, (b) enable personalization of treatment and management options targeted particularly to the specific circumstances and needs of the individual, and (c) help reduce the cost of health care while simultaneously improving outcomes.

History of IoT

1997 - "The Internet of Things" is the seventh in the series of ITU Internet Reports originally launched in 1997 under the title "Challenges to the Network".

1999 - Auto-ID Center founded in MIT

2003 - EPC Global founded in MIT

2005 - Four important technologies of the internet of things was proposed in WSIS conference.

2008 - First international conference of internet of things: The IOT 2008 was held at Zurich.

From any time, any place connectivity for anyone, we will now have connectivity for anything!

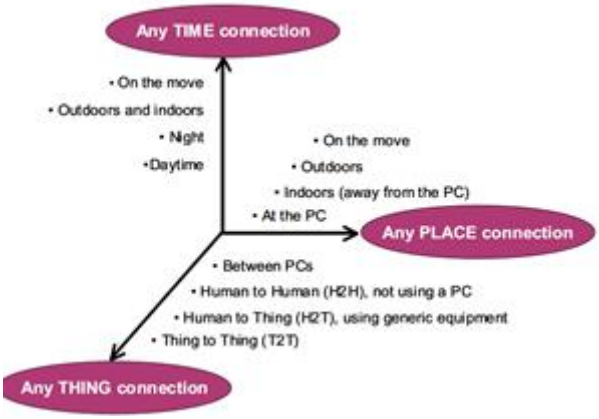
Definitions

As per Wikipedia….The Internet of Things, also called The Internet of Objects, refers to a wireless network between objects, usually the network will be wireless and self-configuring, such as household appliances.

As per Wikipedia…. By embedding short-range mobile transceivers into a wide array of additional gadgets and everyday items, enabling new forms of communication between people and things, and between things themselves.

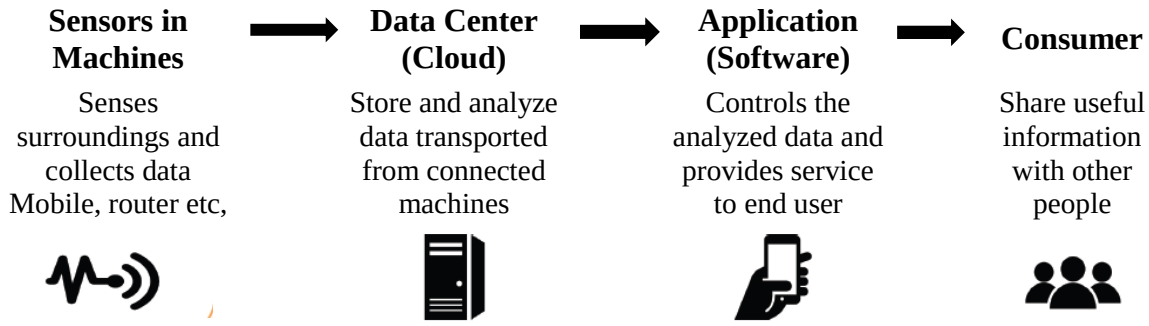
IoT in 2008 …… The term "Internet of Things" has come to describe a number of technologies and research disciplines that enable the Internet to reach out into the real world of physical objects.

IoT in 2020 …… "Things having identities and virtual personalities operating in smart spaces using intelligent interfaces to connect and communicate within social, environmental, and user contexts".



Internet of Things (IoT): Data Flow

Data Flow in IoT is given by the following chart, which illustrates the data transformation from a smart object to the end user consumers



Structure of IoT – Enabling Technologies

The Internet of Things is to connect physical objects over an IP or other network, to exchange/store/collect information to consumers and businesses via a software application. Nearly every person has encountered or used a particular IoT application with 4.9 billion predicted to be connected through 2016. Through this phenomenon, new market opportunities have formed with industries harnessing the IoT potential to further benefit consumers or companies and gain a competitive advantage.

Initially, Radio Frequency Identification (RFID) used to be the governing technology behind IoT development, but with further technological developments, Wireless Sensor Networks (WSN) and Bluetooth enabled devices augmented the mainstream adoption of IoT trend.

Applications of IoT

The ability to network embedded devices with limited CPU, memory and power resources means that IoT finds applications in nearly every field. IoT systems are responsible for performing actions, not just sensing things. Some examples of IoT applications are:*Intelligent shopping systems*, which can monitor specific users' purchasing habits in a store by tracking their specific mobile phones. These users could then be provided with special offers on their favorite products, or even location of items that they need, which their fridge has automatically conveyed to the phone. Applications that deal with heat, electricity and energy management, as well as cruise-assisting transportation systems. Enabling extended home security features and home automation. To describe networks of biological sensors that could use cloud-based analyses to allow users to study DNA or other molecules. With IoT, we can control the electrical devices installed in our house while we are sorting out our files in office. Water will be warm as soon as we get up in the morning for the shower. Entire credit goes to smart devices which make up the smart home.

However, the application of the IoT is not only restricted to these areas. Other specialized use cases of the IoT may also exist. An overview of some of the most prominent application areas is provided here. Based on the application domain, IoT products can be classified broadly into five different categories: smart wearable, smart home, smart city, smart environment, and smart enterprise. There are numerous applications of IoT in many fields. In this paper, we are attempting the IoT applications in Telemedicine, which is gaining importance day by day in healthcare.

IoT in Healthcare of Human beings

IoT devices can be used to enable remote health monitoring and emergency notification systems. These health monitoring devices can range from blood pressure and heart rate monitors to advanced devices capable of monitoring specialized implants, such as pacemakers, Fitbit electronic wristbands or advanced hearing aids. Specialized sensors can also be equipped within living spaces to monitor the health and general well-being of senior citizens, while also ensuring that proper treatment is being administered and assisting people regain lost mobility via therapy as well. More and more end-to-end health monitoring IoT platforms are coming up for antenatal and chronic patients, helping one manage health vitals and recurring medication requirements. IoT in Healthcare is a heterogeneous computing, wirelessly communicating system of apps and devices that connects patients and health providers to diagnose, monitor, track and store vital statistics and medical information. Two important phases of IoT applications in Healthcare

IOT Healthcare solutions can remotely monitor patients suffering from cardiac, diabetes, arrhythmia and chronic diseases, GPS tracking of dementia and Alzheimer's sufferers. Put life-saving data, such as CT scans, test results and patient records, into the hands of medical staff, almost anytime

A) Remote Healthcare-Telemedicine

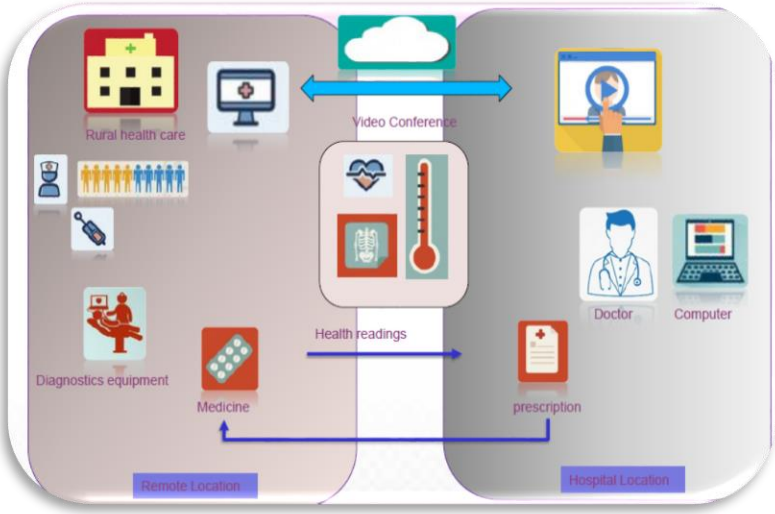
Telemedicine is the use of telecommunication and information technologies in order to provide clinical health care remotely. It helps eliminate distance barriers and can improve access to medical services that would often not be consistently available in distant rural communities. Exchange of information like voice, image, video, graphics, elements of medical record or command to a surgical robot. Features of Telemedicine are

- Remotely diagnose, Record and transfer health parameter of remote patient
- Prescription based diagnostics, Historical record
- Personalized website for patients
- Appointment management-patient can view doctor schedule and can book appointment
- Doctor dashboard, Patient profile and record management
- Inbuilt billing feature
- Bio-sensors to collect vital info –blood pressure, ECG, temperature etc
- Seamless video-conference-multiple camera options, Both real time and offline mode

Telemedicine through a Speech-Based Query System

Patients record questions in their local languages, and doctors follow up through voice or SMS with treatment information. Supports the continuing education of healthcare providers, wherein training materials are created and Multilingual Speech Recognition is used to engage practitioners in a human-like conversation to tests their knowledge and skills

Telemedicine - architecture



Advantages of Telemedicine

- Better reach –extending to access of healthcare to remote area and new markets
- Makes right expertise available anywhere, Reduce unnecessary patient travel
- Improved access and quality of care at comparatively low cost
- Continuity of care, Respond to emergency, Public awareness

B) Outpatient monitoring

- Service for Hospitals –offered to patients on a Hosted model
- Remote Patient monitoring –patients discharged from hospitals
- Heartbeat by heartbeat surveillance, Automated event detection and alerts to Hospital
- 24/7 monitoring of Patients, Real time tracking of outpatient
- Geo-fence, Alerts on deviations, Symptoms recording
- Friends and family alert, Continuous outpatient ECG monitoring
- Cardiac telemetry, Critical message management

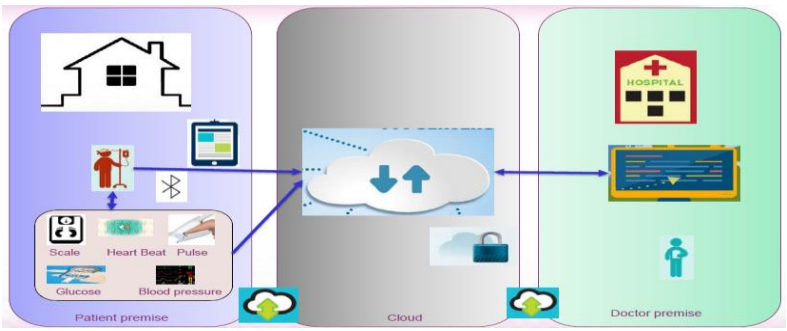
Outpatient monitoring examples

- Chest pain , Mild congestive heart failure
- Hypertension, Weakness, Dizziness
- Vomiting, Altered mental status, Anemia, Mild Asthma
- Back pain, Dehydration

Outpatient monitoring advantages:

- Reduce re-admission –round the clock monitoring for a high risk
- Patients allows the medical practitioner to identify and take corrective actions before grave medical condition.
- Improve efficiency and revenue -round the clock secure access of patient information at a centralized place not only make medical staff more efficient but also allows hospital to effectively manage bed and boost top line
- Can help better patient outcome and overall satisfaction.
- Improve care planning and continued delivery

<u>Outpatient monitoring -architecture</u>



What is Telehealth?

- It is the delivery of healthcare services and clinical information to remote locations
- It is an FDA approved platform that interactively connects patients with a nationwide network of licensed doctors 24/7 using Internet, Internet of Things (IoT), video chats, smartphones and Electronic Medical Record (EMR) clouds
- It is an hour-in-need solution in the 21st century
- It is a new paradigm in the Healthcare industry

Services under TeleHealth Umbrella

- ***TeleMedicine:*** Providing a professional consultation to a patient in a remote location or assisting a primary care physician in rendering a diagnosis. According to the American Medical Association (AMA), 78% of emergency care could be handled efficiently using TeleMedicine
- ***TeleMonitoring:*** Collecting patient data using IoT and sending the data to a healthcare monitoring agency for remote testing and diagnosis. TeleMonitoring services also include personalized alerts that inform a patient's healthcare provider in times of physical/mental trauma
- ***TeleSurgery*****:** Enabling the surgeon to perform an operation on a patient from a distant location using TeleRobotics technology
- ***Remote Medical Education*****:** Providing medical education to the health care service community and targeted groups from a geographically different location
- ***TeleHealth Data Service*****:** Share specialized health information with other Health service providers, the education industry, research firms, and the government etc

Benefits of TeleHealth Services

- Immediate medical attention especially during times of medical emergency and natural disasters
- No need for waiting in long queues to see a physician
- Eliminate the need to physically go to a medical facility. TeleHealth reduces the distance barriers
- Reduced documentation and paperwork
- Cost effective – The growth in TeleHealth space will extensively reduce insurance premiums and potentially reduce the time a patient has to be away from work
- Equal and comprehensive healthcare provisions to everyone by eliminating geographical barriers
- Better communication - Communication to the primary care doctor and specialist happens at the same time because everyone is virtually present in the same room during diagnosis

Few examples of IoT in Healthcare

- Headsets that measure brainwaves, Clothes with sensing devices, BP monitors
- Glucose monitors, ECG monitors, Pulse oximeters
- Sensors embedded in medical equipment, dispensing systems, surgical robots and device implants
- Any wearable technology device……..

CONCLUSIONS:

As discussed in this paper, all the physical objects will work seamlessly with machine to-machine and human-to-machine interfaces. This level of interconnection is a boon for the healthcare, where health influencing factors both internal & external to the human body can be analyzed based on the model. As the examples in this paper make clear, the long predicted IoT revolution in healthcare is already underway. And, as new use cases are emerging, they continue to address the urgent need for affordable, accessible care. Meanwhile, the IoT building blocks of automation and machine-to-machine communication continue to be established. The addition of the service layer forms the complete IoT infrastructure. This revolution is characterized by providing end-to-end processing and connectivity solutions for IoT-driven healthcare. This mobile doctor buddy apps are not meant to be the replacement for experience of the doctors. They should work collaboratively with the doctor. In this approach of complementing the doctor with the technology based inputs, the new trends in IoT has the capability to transform the way the primary healthcare is delivered to the patients. However for the developing world, IoT brings new delivery model for healthcare with good quality at affordable level. It is evident that IoT will facilitate new business models and new healthcare delivery models in the future for both developing and developed worlds, irrespective of the challenges faced at the current time.

In this paper, we reviewed the current state and projected future directions for integration of remote health monitoring technologies into the clinical practice of medicine. Wearable sensors, particularly those equipped with IoT intelligence, offer attractive options for enabling observation and recording of data in home and work environments, over much longer durations than are currently done at office and laboratory visits. This treasure trove of data, when analyzed and presented to physicians in easy-to-assimilate visualizations has the potential for radically improving healthcare and reducing costs.

REFERENCES:

[1]. Gartner, IT Glossary, Internet of Things - http://www.gartner.com/it-glossary/internetof-Things

[2]. Internet of Things (IoT): A Literature Review Somayya Madakam et al, Jou of Comp & Comm. 2015, 3, 164-173

[3] Lianos, M. et al (2000) Dangerization and the End of Deviance: The Institutional Environment. British Journal of Criminology, 40, 261-278

[4]. Aggarwal, R. et al (2012) RFID Security in the Context of "Internet of Things". First International Conference on Security of Internet of Things, Kerala, 17-19 August 2012, 51-56. http://dx.doi.org/10.1145/2490428.2490435

[5]. Gigli, M. et al (2011) Internet of Things, Services and Applications Categorization. Advances in Internet of Things, 1, 27-31. http://dx.doi.org/10.4236/ait.2011.12004

[6]. (2005) ITU Internet Reports, International Telecommunication Union. The Internet of Things: 7th Edition. www.itu.int/internetofthings/on

[7]. *"Internet of Things Global Standards Initiative"*. *ITU*. Retrieved 26 June 2015.

Prescriptive Analytics for Intelligent business systems

Raghuvira Pratap A[1],
V.R.SIDDHARTHA ENGINEERING COLLEGE
[1]raghuvirapratap@gmail.com

J V D Prasad[2]
V.R.SIDDHARTHA ENGINEERING COLLEGE
[2]prasadjasti@yahoo.co.in

Kranthi Kumar G[3],
V.R.SIDDHARTHA ENGINEERING COLLEGE
[3]gkk15@rediffmail.com

Dr. Suvarna Vani K[4]
V.R.SIDDHARTHA ENGINEERING COLLEGE
[4]suvarnavanik@gmail.com

***Abstract*— Data is vantage and valuable. Across world, "big data" and analytics are helping businesses to become smarter, more productive, and better at making predictions. The internet has engendered an explosion in data growth in the form of data sets, called Big Data, that are so large they are difficult to store, manage and analyze using traditional RDBMS which are tuned for Online Transaction Processing (OLTP) only. Big Data is changing the way analytics were commonly viewed, from data mining to Advanced Analytics. Big Data is composed of text, image, video, audio, and mobile or other forms of data collected from multiple datasets, and is rapidly growing in size and complexity. It has created a huge volume of multidimensional data within a very short time period. A real-world application could be modelled as a multi objective, dynamic, large scale optimization problem. It is recognized that the prescriptive analytics based techniques are good ways to handle this kind of problems. Based on the utilization of business intelligence systems, the real-world system will be more efficient and effective. We crystallize the availability of new in-memory technology high-performance analytics and prescriptive analytics that works on business intelligence systems in providing a better way to analyze data more quickly than ever.**

***Keywords*— big data analytics, unstructured data analysis, prescriptive analytics; business intelligence systems**

I. INTRODUCTION

Big data is creating unsurpassed opportunities for businesses to achieve deeper, faster insights that can strengthen decision making, improve the customer experience, and accelerate the pace of innovation. Basically, "Big Data" science doesn't mean only a large volume of data but also other features that differentiate it from the concepts of "enormous data" and "Huge data" [1][2]. But today, most big data yields neither meaning nor value. Businesses are
so overwhelmed by the amount and variety of data cascading into and through their operations that they
struggle just to store the data—much less analyse, interpret, and present it in meaningful ways [2].

Analytics became a competitive change in a big data world that is rapidly transmuting into the digital business epoch. The huge amounts, difficulty and variation of data sources in our data-driven economy are already shattering existing self-service Business Intelligence data discovery tools that can't render the data. The human analytics or decision maker's ability to manually identify new perceptions or detect changing patterns efficiently with those tools when presented with hundreds of variables is also being surpassed. Most decisions can't wait for sparse data scientist talent to evaluate prospective recourse. As a result, the business ends up operating on riskier gut feel rather than data-driven decisions [4].

Business intelligence & analytics (BI&A) has evolved to become a foundational cornerstone of enterprise decision support. Quick-witted intelligence data discovery tools deliver simple advanced analytics services that are designed especially for the non-data expert, information operative. Unlike the current self-service BI/data discovery tools that include limited forecasting, trend lines, outlier highlighting and other basic insight detection, smart data discovery tools provide intelligent data preparation and much deeper, statistically significant, guided diagnostic exploration. Another key difference is that business users can experiment with what-if scenarios and get unbiased, detailed written interpretations of prescriptive actionable insights [5].

Self-service business intelligence (BI), enabling a throng of users to easily mash up data from

a wide range of sources— click streams, social media, log files, videos, and more. With the aid of high-configured desktops and mobile computing devices, users can perform real-time, predictive analyses, and showcase the results in compelling, interactive, and easily understood visual formats. The trend strive towards visualization-based data discovery tools is worth exploring by any business that seeks to derive more value from big data. Government agencies and large corporations are launching research programs to address big data's challenges. Visualization in today's time is very effective for presenting essential information in vast amounts of data.

Prescriptive data analytics addresses information obtained through comment, measurement, or tests about a phenomenon of interest.
The following lists only a few potential purposes:

a) To generalize and deduce the data and determine how to use it.
b) To check whether the data are genuine.
c) To give guidance and contribution in decision making system.
d) To identify and conclude reasons for fault.
e) To forecast what will occur in the future.

Descriptive Analytics: makes use of historical data to examine what occurred in past. For instance, a reversion technique may be used to find simple trends in the datasets, visualization presents data in a meaningful fashion, and data modelling is used to collect, store and cut the data in an efficient way. Descriptive analytics is typically associated with business intelligence or visibility systems.
Predictive Analytics: emphasis on predicting future probabilities and trends. For example, predictive modelling uses statistical techniques such as linear and logistic regression to understand trends and predict future out-comes, and data mining extracts patterns to provide insight and forecasts.
Prescriptive Analytics: addresses decision making and efficiency. For example, simulation is used to analyse complex systems to gain insight into system performance and identify issues and optimization techniques are used to find best solutions under given constraints.

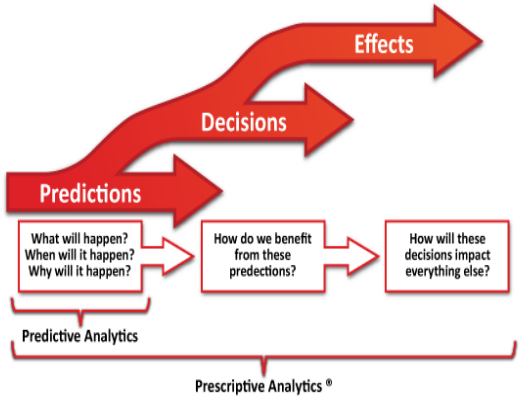


Figure 1. Prescriptive Analytics extends beyond predictive analytics by specifying both the actions necessary to achieve predicted outcomes, and the interrelated effects of each decision.

Prescriptive analytics is also concerned with the future – but instead of just predicting what will happen without intervention, this takes the analysis one pace further. Decision-based analytics, which is another term for prescriptive techniques, allows organizations to act on the data being analyzed, adapt quickly and better serve their markets, all the way down to individual customers. Despite being so new, prescriptive analytics promises to be extremely powerful and accurate in its predictions. Prescriptive algorithms use a large variety of techniques, such as machine learning, artificial intelligence, and mathematical sciences, to understand the impact of future decisions and adjust actual decisions based on that outcome. This will drastically improve decision making as it incorporates future possible outcomes when making a prediction[6].

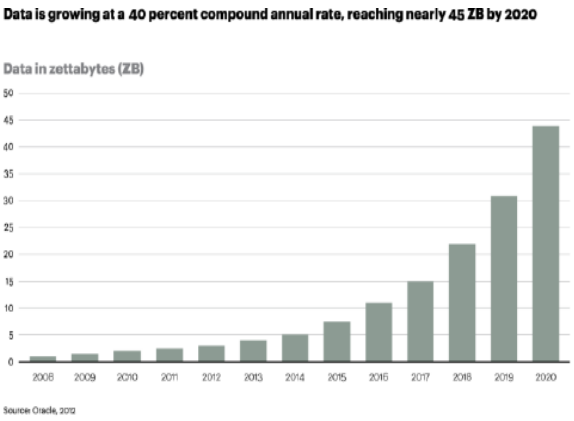


Figure 2. Growth of Big-data

Prescription in today's time is very effective for presenting essential information in vast amounts of data. Big-data discovery tools present new research opportunities to the graphics and prescriptive community [3]. The size of the collected data about the Web and mobile device users is even greater. To provide the ability to make sense and maximize utilization of such vast amounts of data for knowledge discovery and decision making is crucial to scientific advancement; we need new tools beyond conventional data mining and statistical analysis. Prescriptive analytics is a tool which is shown to be effective for gleaning insight in big data.

We need to focus on technologies that are emerging to support back-end concerns such as storage and processing. Prescriptive-based data discovery tools focus on the front end of big data-on helping businesses explore the data more easily and understand it more fully.

II. BACKGROUND STUDY

Analytics is the discovery of meaningful patterns in data. While it is not new, we're going through a renaissance in data science and technology in business analytics. What started as Descriptive

Analytics, using data to build reports and dashboards to present it in a more consumable way, grew to Investigative Analytics to answer the 'why' question, has matured into Predictive Analytics, with historical data, models and algorithms to predict the future outcomes. The next leap is Prescriptive Analytics which not only predicts the outcomes but also suggest or prescribe a solution in order to influence the future in the desired way [6].

Prescriptive analytics synergistically combines data, business rules, and mathematical models. The data inputs to prescriptive analytics may come from multiple sources, internal (inside the organization) and external (social media, et al.). The data may also be structured, which includes numerical and categorical data, as well as unstructured data, such as text, images, audio, and video data. Business rules define the business process and include constraints, preferences, policies, best practices and boundaries. Mathematical models are techniques derived from mathematical sciences and related disciplines including applied statistics, machine learning, operations research, and natural language processing[7].

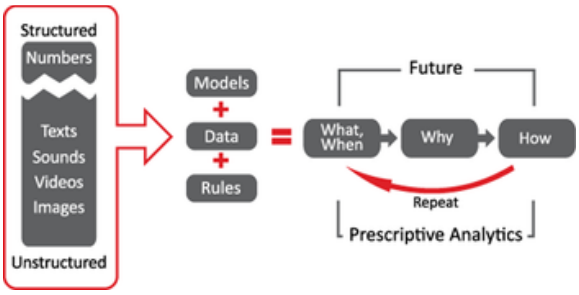


Figure 3. Prescriptive Analytics process.

Prescriptive analytics are comparatively complex in nature and many companies are not yet using them in day-to-day business activities, as it becomes difficult to manage. Prescriptive analytics if implemented properly can have a major impact on business growth. Large scale organizations use prescriptive analytics for scheduling the inventory in the supply chain, optimizing production, etc. to optimize customer experience.

Prescriptive Analytics Methodology

Through a variety of statistical modeling approaches, Prescriptive Analytics helps businesses predict the behavior of key variables that are unknown, yet have significant impact on the performance of the business. Prescriptive models are also used for analyzing information patterns to support tactical analytics, such as fraud detection or online marketing.

Prescriptive Analytics translates a forecast into a feasible plan for the business, and helps users identify the best steps to implement. There are two primary approaches – simulation and optimization.

A: Simulation is best used in design situations, where it helps users identify system behaviors under different configurations, and ensures all key performance metrics are met (e.g. wait times, queue length, etc.).

Explanation of Simulation Steps

- ✓ Step1: **Understanding the business:** An important first step to realizing the potential benefits of big data for business is deciding what the business model(s) will be the data economy supports an entire ecosystem of businesses and other stakeholder organizations. These are often dependent upon each other's products and services so the vitality of the sector as a whole is crucial.

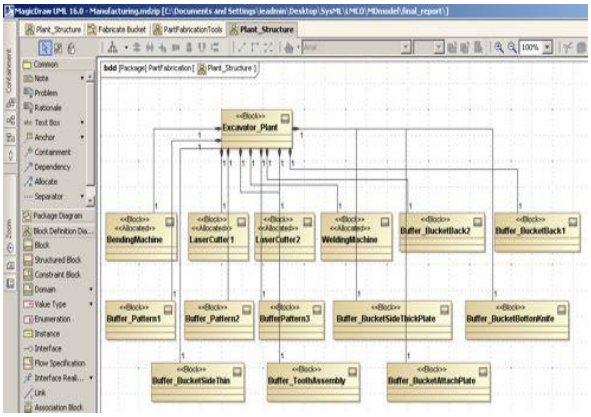


Figure 4: Simulation Generator to understand business and data

- ✓ Step 2: **Understanding the data:** Big data businesses can essentially be categorized as data users, data suppliers, and data facilitators. These are not mutually exclusive and as many firms engage in a range of activities. Data users are organizations that use data internally—either for business intelligence activities such as forecasting demand, or as an input into other products and services such as credit scores or targeted advertising.
- ✓ Step 3: **Data preparation/cleaning:** They either generate data internally or acquire it from third parties (or both). The key question for this group is "what data do we have and how can this data be used to create value within the business?"

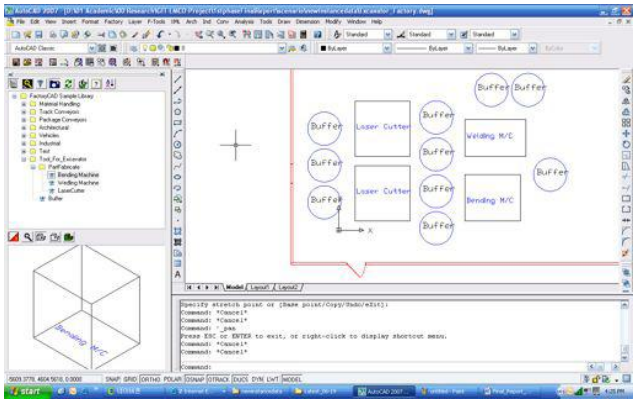


Figure 5: Predictive Approach to understand business and data

✓ Step 4**:Creating a predictive model:** Predictive modeling uses statistics to predict outcomes.Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred. For example, predictive models are often used to detect crimes and identify suspects, after the crime has taken place.

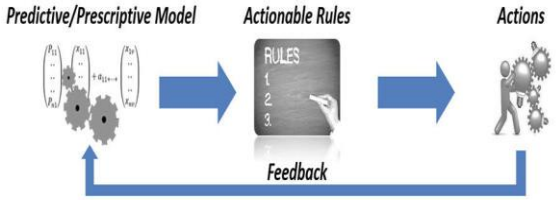


Figure 6. Prescriptive Analytics Model.

B: Optimization supports ongoing operational, tactical and strategic business planning; it leverages linear programming to identify the best outcome for a business, given constraints and objective function.

Simulation-based optimization integrates optimization techniques into simulation analysis. Because of the complexity of the simulation, the objective function may become difficult and expensive to evaluate.

Explanation of Optimization in prescriptive analytics

Once a system is mathematically modeled, computer-based simulations provide the information about its behavior. Parametric simulation methods can be used to improve the performance of a system. In this method, the input of each variable is varied with other parameters remaining constant and the effect on the design objective is observed. This is a time-consuming method and improves the performance partially. To obtain the optimal solution with minimum computation and time, the problem is solved iteratively where in each iteration the solution moves closer to the optimum solution. Such methods are known as 'numerical optimization' or 'simulation-based optimization'.

Specific simulation based optimization methods can be chosen based on the decision variable types. Optimization exists in two main branches of operational research

Optimization parametric and Optimization control. Here Optimization Parametric is a static process to find the values of parameters "static" for all states, with the goal of maximize or minimize a function. In this case, there is the use of mathematical programming, such as linear programming. In this scenario, simulation helps when the parameters contain noise or the evaluation of the problem would demand excess of computer time, due to its complexity

Optimization control (dynamic) – used largely in computer sciences and electrical engineering, what results in many papers and projects in these fields. The optimal control is per state and the results change in each of them. There is use of mathematical programming, as well as dynamic programming. In this scenario, simulation can generate random samples and solve complex and large-scale problems.

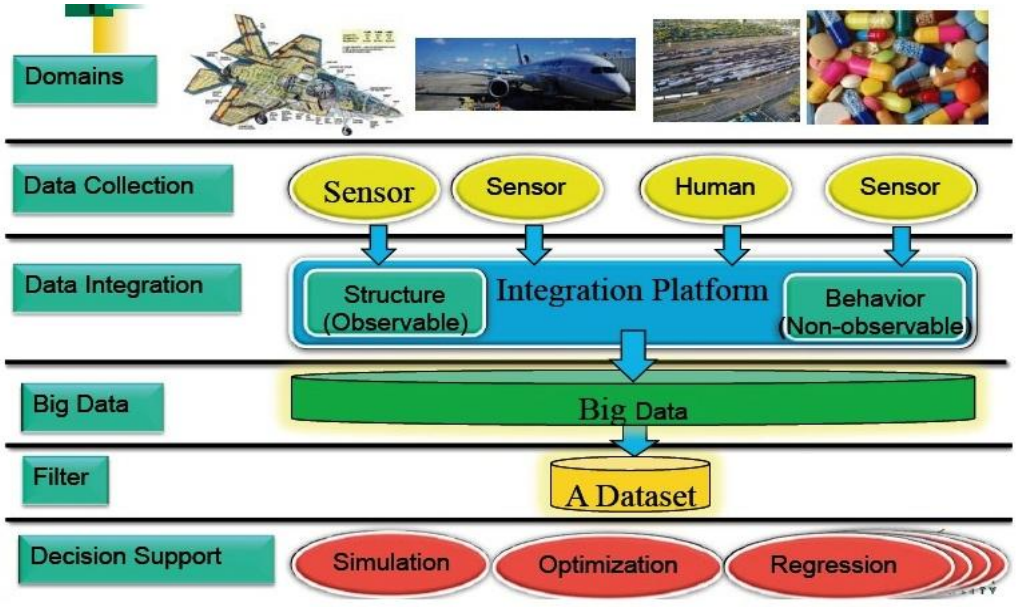


Figure 7. Different stages of Prescriptive Analytics Process

One of the main approach in Simulation Optimizations is Statistical ranking and selection methods (R/S). In Statistical ranking and selection method it is designed for problems where the alternatives are fixed and known and simulation is used to estimate the system performance. In the simulation optimization setting, applicable methods include indifference zone approaches, optimal computing budget allocation, and knowledge gradient algorithms.

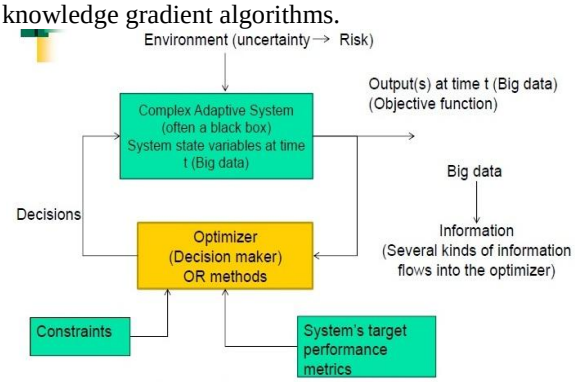


Figure 8: Simulation optimization process for decision making

III.PROPOSED APPROACH

Prescriptive analytics approach for big data decision making in business intelligence system is to analyse a huge or even distributed dataset, several analysis models usually serve different purposes and provide unique insights.

- Each modelling technique might be capable of answering specific questions or only partial data set.
- Complex problems require multiple models interoperating to complement/supplement each other.

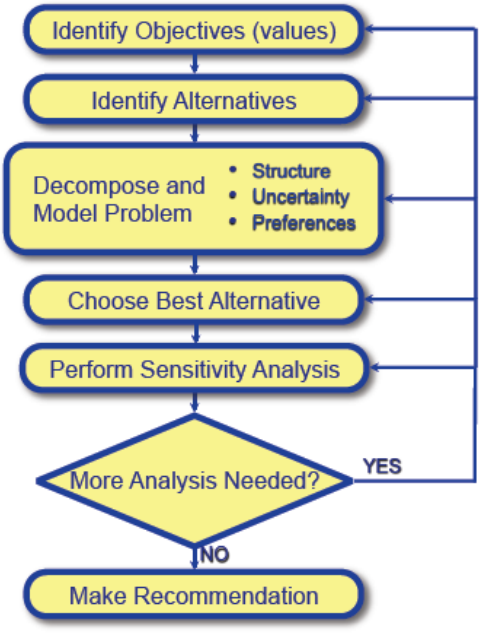


Figure 8: Prescriptive analytics approach

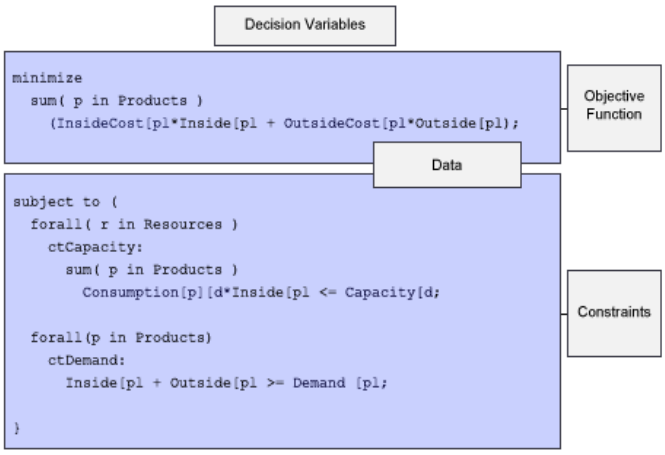
New Approach of Prescriptive Analytics

Step 1 Optimization at segment level:

- ✓ Easily visualize segments and actions using a familiar Decision Tree view
- ✓ Assign automatic treatments through optimization and rule-based decisions

Linear Optimization Algorithm: Linear programming (LP) (also called linear optimization) is a method to achieve the best outcome (such as maximum profit or lowest cost) in a mathematical model whose requirements are represented by linear relationships. The concept behind a linear programming problem is simple. It consists for four basic components:

1. Decision variables represent quantities to be determined
2. Objective function represents how the decision variables affect the cost or value to be optimized (minimized or maximized)
3. Constraints represent how the decision variables use resources, which are available in limited quantities
4. Data quantifies the relationships represented in the objective function and the constraints



Step 2 Simulation based Optimization methodology

1. Utilize proprietary optimization methods to map the cost structure.
2. Determine appropriate constraints, costs and revenue per unit
3. Validate the model through back-testing to actual financial statement.
4. Implement the optimization strategy
5. Monitor the progress of attaining the optimal strategy and revise on quarterly base.

Step 3 Regression based optimization

In Regression approach, the aim is to capture the interdependencies between outcome variables and explanatory variables, and exploit them to make predictions. Regression technique addresses continuous outcome variables (eg. business values). A common assumption is exogenetic assumption. The validity of most statistical methods used in

regression analysis depends on this assumption. A Regression based models the past relationship variables to predict the future behaviour.

There are several different classes of regression procedures, with each having varying degrees of complexity and explanatory power. The most basic type of regression is that of simple linear regression. A simple linear regression uses only one independent variable, and it describes the relationship between the independent variable and dependent variable as a straight line.

Equation of a Regression Line:

$$f(x) = mx + b$$

Variables, constants, and coefficients are represented in the equation of a line as

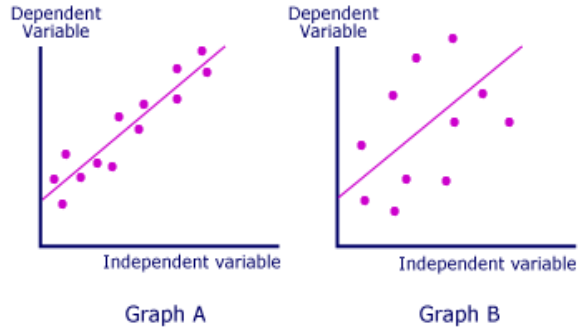
x represents the independent variable
f(x) represents the dependent variable
the constant b denotes the y-intercept—this will be the value of the dependent variable if the independent variable is equal to zero
the coefficient m describes the movement in the dependent variable as a result of a given movement in the independent variable

IV. APPLICATIONS AND RESULTS

For Regression based optimization

The placement office of a graduate business school would like to predict the starting salaries of its students. The placement office administrators are highly confident (based on their collective past experience) that starting salaries depend on a combination of factors; including the number of years of previous work experience, the student's graduate school CGPA, and the student's CAT score. Is it appropriate for the placement office to use a simple linear regression to predict the starting salaries of its students?

The following two graphs illustrate simple linear regressions. Which has a higher predictive quality?



For Optimization at segment level consider as an example, imagine that your company wants to understand how past advertising expenditures have related to sales in order to make future decisions about advertising. The following table lists the monthly sales and advertising expenditures for all of last year by a digital electronics company.

Month	Sales (in 1000s)	Advertising Dollars (1000s)
January	100	5.5
February	110	5.8
March	112	6
April	115	5.9
May	117	6.2
June	116	6.3
July	118	6.5
August	120	6.6
September	121	6.4
October	120	6.5
November	117	6.7
December	123	6.8

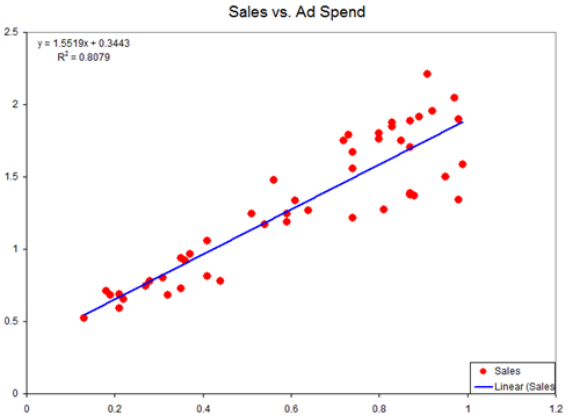


Figure 9: Optimization Process. We can make an intuitive assessment that increases in Ad spend also icreases the sales. Using the straight line, we may also be able to predict.

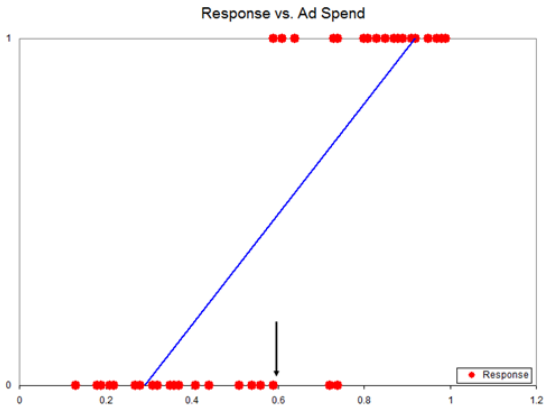
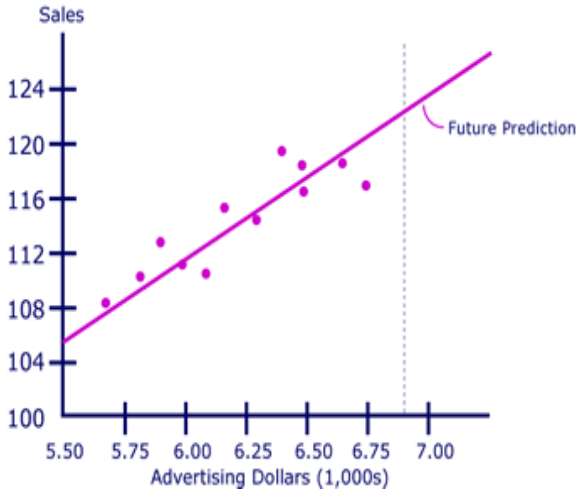


Figure 10: Linear fit for binary outcome: Although we can make ann intuitive assessment that increase in Ad

spend increases Response , the switch is abrupt around 0.6



The extension of the line of regression requires the assumption that the underlying process causing the relationship between the two variables is valid beyond the range of the sample data.

V.CONCLUSION

With the amount of data growing constantly and exponentially, the data processing tasks have been beyond the computing ability of traditional computational models. To handle these massive data, i.e., deal with the big data analytics problem, more effective and efficient methods should be designed. In business intelligence system and evolution Identifying and making decisions about opportunities areas of unmet need, predicting the potential upside , Proactively tracking industry needs and trends, Exploiting data analytics to identify specific consumer populations, Leveraging data analytics to identify key innovations for product optimization that will generate the largest investment return is much needed. This paper has reviewed the connection between data analytics and business intelligence/evolutionary systems. The potential combination of data science and business intelligence in optimization and data analytics was also analysed. Data science involves prediction or inference on a large amount of data. Swarm intelligence studies the collective behaviours in a group of individuals. The Systems are not running at optimal level of performance and there is scope for improvement. Here data generated by the business intelligence system must be extracted, visualized. Statistically analysed, and converted into information. Decision making is the optimization of decision parameters to meet the objectives of the business systems that is subject to constrains and real world decision making happens under uncertainty. Finally it is necessary that the whole prescriptive analytics effort is business driven, with a good understanding of where the major payoffs are and how decisions should be prioritized. For organizations inexperienced in the domain this may mean using external resources to formulate a strategy.

VI.REFERENCES

[1] C Gröger, H Schwarz, B Mitschang " Prescriptive analytics for recommendation-based business process optimization"

[2] Christoph Gr , Holger Schwarz and Bernhard Mitschang , " Prescriptive Analytics for Recommendations- Based Business Process Optimization " ,Business Information Systems Volume 176 of the series Lecture Notes in Business Information Processing pp 25-37.

[3] Muehlen, M.z., Shapiro, R.: Business Process Analytics. In: Vom Brocke, J., Rosemann, M.(eds.) Handbook on Business Process Management 2, pp. 137–158. Springer, Berlin (2010)

[4] Kemper, H.-G., Baars, H., Lasi, H.: An Integrated Business Intelligence Framework. Closing the Gap Between IT Support for Management and for Production. In: Rausch, P., Sheta, A.F., Ayesh, A. (eds.) Business Intelligence and Performance Management, pp. 13–26. Springer, London (2013)

[5] Radeschütz, S., Mitschang, B., Leymann, F.: Matching of Process Data and Operational Data for a Deep Business Analysis. In: Interoperability for Enterprise Software and Applications (IESA), pp. 171–182. Springer, Berlin (2008)

[6] Groger, C., Schlaudraff, J., Niedermann, F., Mitschang, B.: Warehousing Manufacturing Data. A Holistic Process Warehouse for Advanced Manufacturing Analytics. In: Data Warehousing and Knowledge Discovery (DaWaK), pp. 142–155. Springer, Berlin (2012).

[7] Riza, Bergmeir, Herrera, Benìtez, 2015 , Fuzzy Rule-Based Systems for Classification and Regression in R - Journal of Statistical Software, Vol 65, Issue 6.

[8] Dan Sommer, Rita L. Sallam, James Richardson, "Big Data Visualization:Turning Big Data Into Big Insights" March 2013 , Intel IT Center

[9] Erlach, K.: Value stream design. The way to lean factory. Springer, Berlin (2011)

[10] Ng R.T. and Han J. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining, Proc. 20th Int. Conf. on Very Large Data Bases, 144-155. Santiago, Chile.
J. MacMillan and S. B. Van Hemel (eds.), Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education, Washington, DC: The National Academies Press.

[11] Fitriana R, Eriyatno TD. Progress in Business Intelligence System research: A literature Review. International Journal

[12] O'Reilly, T. 2005. "What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software," September 30, (http://www.oreillynet.com0

[13] Malladi S, editor Adoption of Business Intelligence & Analytics in Organizations–An Empirical Study of Antecedents.19th American Conference on Information Systems (AMCIS); 2013; Chicago, Illinois.

Implementing Power Distribution System Using Geographic Information System

A. Sowmya[1], A. Jitendra[2]

[1]PG Scholar, Dept. of CSE, V.R Siddhartha Engineering College,

[2]Assistant Professor, Dept. of CSE, V.R Siddhartha Engineering College

[1]E-mail:somuankam@gmail.com

[2]E-mail:jitendra@vrsiddhartha.ac.in

Abstract: **Electricity is an essential need for our daily life. Power distribution should be managed safely and effectively by power distribution companies. Efficient functioning of Distribution Company leads to the development of power sector and economy. Power distribution system is benefited much more with the advancement of software and IT sector. This paper shows the application of new and emerging technology like GIS which plays a major role in modern management of power distribution companies. GIS works with data on an interactive map where it can be updated, understood, and shared. GIS integrates both land base and the electrical network maps. GIS is not only useful in improving internal efficiency levels pertaining to power supply monitoring, developing accurate database, commercial and customer services but also extremely useful for important functions like facility management, energy audit, network analysis, trouble call management, load management, theft detection etc.**

I. INTRODUCTION

Distribution sector in India is the weakest area compared to Generation & Transmission. Most of the losses and faults occur in this particular area which directly affects the consumer. The situation is worse in some rural areas due to insufficient attention in transmission & distribution losses. The average T&D losses are currently at 28% by the year 2012. Some common reasons for the high losses are lack of proper administration & power theft. Thus the present distribution system requires to be upgraded to minimize the loss. By applying GIS we can reduce the distribution planning cost by reducing the material cost through Optimum feeder path. Economic importance of distribution system is very high. The amount of investment involved dictates careful planning, design, construction and operation which assure growing demand for electricity in terms of growing rates and high load densities.

GIS, a solution to these problems, is a graphical advanced version of SCADA in which electrical(technical) parameters of system are displayed on graphical map with respect to its physical locations (point of connection). Social commercial environmental effects become part of same information system. In normal display system only technical parameters are given importance whereas in GIS, collective approach for parameters as well as its effects on surroundings is followed.

II. PROBLEMS FACED BY DISTRIBUTION SYSTEM

Growth of automation and use of modern equipments demand better quality of power. Hence with the power distribution, power quality has also been taken into consideration by the electric utilities. There are various problems faced by utilities which are:

A. Increased Equipment Loading:

Most utilities have increased levels of "asset utilization" due to short-term financial pressures. High equipment loading is well understood from the perspective of thermal aging and conductor sag. With everything else equal, high loading increases failure probability. Detailed failure rate models do not exist. But the probability of second-order failures increases with the square of failure rate. The probability of third-order failures increases with the cube of failure rate, and so forth. Aside reliability, thermal aging of organic insulation increases exponentially with temperature. This substantially impact the useful life of moderately loaded equipment, but becomes a financial concern when systematic increase in equipment loading begin to materially reduce useful life.

B. Ageing Infrastructure:

Electricity usage grew at an annual rate of approximately 7% before the 1970s. This implied that 14% of equipment would have been older than 30 years and 0.5% would have been exceeding 50 years without considering failures. Growth has been lower at approximately 2.5%, resulting in minimal procurement need for new equipment for the last 30 years. This implies currently 8% of existing equipment is older than 50 years and 49% older than 30 years. Aging infrastructure is a major problem due to growth rate alone, is increased by higher equipment loadings and less aggressive replacement programs, and has been recognized by the Department of Energy and one of the major issues facing electric utilities.

C. Increased demand for reliability and power quality:

Many customers are demanding higher levels of power quality and reliability, while utilities are under increasing pressure to reduce cost and deal with aging infrastructures. Long interruptions halt Production. Short interruptions cause computer systems to crash. Waveform distortions, such as sags, can cause motor contacts to drop out and electronic controls to malfunction. Many customers are not willing to pay for increased quality. Many see perfect reliability as an entitlement and as an opportunity to ride free on others who are willing to pay for premium service. Different customers have different needs and existing distribution systems are not able to differentiate reliability accordingly. Reliability is too high for most, too low for some, and just right for few.

D. Some other problems faced by distribution system are:

1) Large gap between supply and demand due to shortage of power.
2) Employees neglect the problems of consumers due to monopoly in power sector.
3) Most of the substations and transformers are affected by overloading.
4) Revenue collection is poor due to which financial losses are higher.
5) Power quality issues such as interruptions, flickers and poor voltage.
6) Modernization is not possible due to less investment in power sectors.
7) Transmission & Distribution losses are more which are up to 45%.
8) Lack of effective management, control and proper communication.

III. EXISTING DISTRIBUTION SYSTEM

Generation, transmission and distribution are the main parts of electrical power system. The structure of transmission and distribution system covers a huge network with of a wide range of equipment, feeders and services. Each system has its own role. Typical systems used in the electrical utility are SCADA, DMS, NA and ERP. SCADA is a complex network of electronic measuring and sensing instruments for capturing the data which is then communicated over LAN and WAN to the control center. SCADA carries the function of monitoring the utility network and provides the remote control of switching devices, transformers and equipments. This facilitates utilities to carry out the maintenance and fault rectification activities of the distribution system. DMS (Distribution Management System) supports operational improvements by using online network and it is also used to efficiently manage the 11KV and below network by providing planned switching orders and load flow analysis to minimize losses and equipment overloads. Traditional SCADA systems are early smart grid technologies. But the use of SCADA is limited to a few substations and major distribution automation devices. The data management by SCADA plays an important role in any smart grid implementation. Combination of smart meters, data management, communication network and applications specific to metering is advanced metering infrastructure (AMI). AMI plays a key role in smart grid technology and many utilities begin smart grid implementation with AMI. GIS is used to superimpose the complete electrical network assets from generation to service point on top of the land base data.

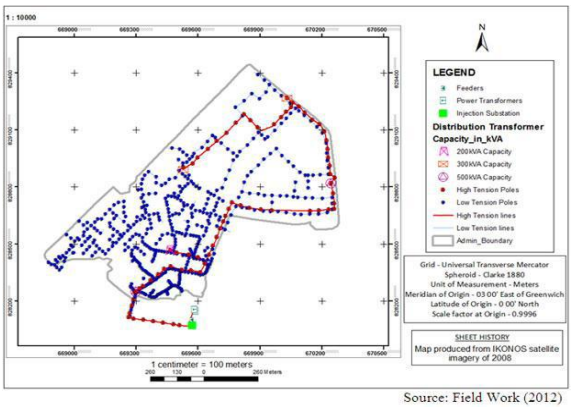
IV. GEOGRAPHIC INFORMATION SYSTEM (GIS) IN POWER DISTRIBUTION SYSTEM

GIS technology is widely implemented in energy sector, especially with the advancement in modern management systems and automation. GIS provides good platform for system representation and manipulation, since network models and data bases can be easily accessed and modified to perform system analysis. GIS plays a strong role in the management of distribution system. It is a framework that manages an electric utility information technology system. GIS creates spatial information about utility assets (poles, wires, transformers, duct banks, customers) and serves that information to the

utility. The combined data served from the GIS and SCADA is combined along with other information from outside the utility such as weather systems, traffic, or satellite imagery.

This combined information is used by utilities to visualize a common operating picture for monitoring and maintaining network analysis and planning. Relationships between systems and the environment can be understood with the application of GIS. GIS can show the view of the grid and note the changes. GIS when compared with a SCADA system can show the complete state of the network represented by a realistic model. Fault location can be identified easily and quickly by combining GIS and fault passage indicator (FPI).

Using the digitization facilities of the software the automated mapping(AM), helps the utilities to quickly create digital maps of their supply area. The maps when digitized contain its precised location, technical information and detailed information about the site service by the utility and of the distribution network equipment that are installed in the field. For example, when an employee wants to know the date of installation of a given transformer, he just clicks on that transformer symbol. The represented attributes of that transformer will show him the installation date along with the information related to it. Assume that the same employee now further wants to know more complex information. For example if he wants to see only 100 KVA transformers on the map installed prior to a given date, the query facilities the software to quickly process his requirement and show on the map only those transformers of 100 KVA, hiding all other transformers. Similarly if he wants to assess the requirement of a cable to be laid along a certain road, the GIS will return him the results of processing considering even all the bends and turns the road. The cable length so shown by the GIS will be precise and will therefore help him procure the exact required quantity of the cable.



Source: Field Work (2012)

Fig 1: Electricity Distribution Network Map

V. CASE STUDY FOR THE APPLICATION OF GIS

Roorkee which belongs to the circle of Uttaranchal Power Corporation Ltd (UPCL), has a consumer base of 1.2 lacks. GIS technology has been effectively implemented in Roorkee. Following activities are done in implementing GIS:

A. Meter Installation Survey:

Meter installation survey is done to find out the condition of the metering equipment at the consumer's premises. Also consumer details are collected and updated in the database using GIS tools.

B. Network Mapping:

The location coordinates (Latitude-Longitude) of every consumer is plotted in GIS map. Electrical network element, from 33 KV sub-station through 11 KV feeder down to DT and the nearest LT service are plotted on GIS map, with the following features:

1) All the network elements are identified. A database is developed to record all the technical attributes of the network element.

2) All the *network* assets are given with a unique identification number. The network database has a linkage with consumer database.

3) The network database has a GUI interface in which all the child components are shown as subset of the parent. When a parent is selected the entire child components can be seen in the left pane. The graphical symbol of the parent component is shown as expandable.

4) In case of network reconfiguration where some components are electrically connected to a new parent component, then all such child components can be selected in the left pane can be dragged and dropped to be new parent component. The database gets immediately modified to show new electrical connectivity.

5) The entire electrical network has been mapped on a scale of 1:4000. When the DT is selected on GIS map, then all the LT lines connected to that DT and corresponding LT lines are shown.

C. Consumer Indexing:

A unique Consumer Index Number (CIN) is given to all types of consumers and the consumer-network database has been developed for correlating each consumer to the corresponding electrical attributes, using GIS tools to query and retrieve

information. The methodology adopted for the exercise has been given below:

1) A detailed door-to-door consumer survey was carried out for the creation of consumer database linked to DT (for LT customers) and linked to feeder for HT consumers.

2) All consumers were allotted a unique Consumer Identification Number (CIN) based on the electrical address of consumers.

3) The information of the consumer's network connectivity has been maintained in the database.

4) The consumer database has been linked to the network database for the purpose of defining the consumer's electrical connectivity.

D. Distribution Network Modeling:

The electrical assets like Sub-stations, 11 KV feeders, DT's, Poles and LT feeders have also been uniquely codified and modeled with the help of GIS and GPS technologies.

E. Load Flow Analysis:

This is being done with the help of electrical database imported from GIS map. It plays a key role in determining technical loss, planning and optimization of distribution system.

VI. CONCLUSION

GIS plays an important role in establishing communication between automation systems like SCADA, DMS, AMR and customer care and billing systems. GIS technology helps in fast, accurate and reliable data management. Since the sub-transmission and the distribution network of a power utility have a Geographical reference, it is beneficial to create the network on GIS map and constantly update the same as per field parameters. With periodic updating and monitoring, GIS mapping of the Electrical Network and Consumer database helps in improved load management, loss reduction, better revenue realization, asset and work management and possibly better consumer relationship.

VII. REFERENCES

1. Anil Kumar, Arun M Shandilya, S. K. Katiyar, "Application of GIS And GPS in Distribution Power Sector", IJAEE, 2013.
2. Nidhi Mishra, Kranti Suresh Khair, Priyanka Parikshit Pawar , Poonam Baban Thakur, Pooja Tanaji Satpute, "Management Of Distribution System Using GIS", IJERA, 2014.
3. Dr. Tripta Thakur, "GIS Based Power Distribution System: A Case study for the Bhopal City".
4. Surendra Kumar Yadav, "GIS in Power Sector Management", IJERT, 2013.
5. N. V. Vader, Mrs. S. S. Kulkarni, "Improving Efficiency of power sectors by using GIS", National Conference on Geo – Informatics, Thane.
6. N.Vijyee, "Mapping Of Chennai Electricity Distribution Network", 10th ESRI India User Conference, 2009.
7. Nagaraja Sekhar, K.S.Rajan, Amit Jain, "Application of Geographical Information System and Spatial Informatics to Electric Power Systems", NPSC, 2008.
8. Jayant Sinha, "GIS application in Power Distribution Utility", UPCL, Dehradun.

A Study on Social Engineering Attacks and Defence Mechanisms

Mukesh Chinta [#1], Jitendra Alaparthi [#2], Eswar Kodali [#3]

[#1 , #2, #3] *Department of CSE, V R Siddhartha Engineering College*
Vijayawada, Andhra Pradesh, India

[1] mukesh.chinta@vrsiddhartha.ac.in
[2] jitendra@vrsiddhartha.ac.in
[3] kodalieswar77@gmail.com

Abstract—**Humans are the most vulnerable points in any kind of security system because of their predictable behaviour and other psychological aspects. Yet, a lot of emphasis related to security is given to implementation of technical security via an antivirus, Intrusion Detection System (IDS), Intrusion Prevention System (IPS), Firewalls etc ignoring the nontechnical behaviour altogether. This is where, Social Engineering- concept of exploiting computer systems and individuals alike has become a major concern not just for organization but also for common people. This paper introduces the concept of social engineering, different types, common ways of attack and related case studies. In addition, several ways to defend against social engineering by proper education, training, procedures and policies are also discussed. Ultimately, this paper highlights the fact that social engineering has grown to be one of the potent threats to information security and should be given equal importance to its technological counterparts.**

I. Introduction

Information security has always been a major concern for organizations over the past decade. Most of the techniques developed to counter the security attacks concentrated on putting up technical barricades around the systems trying to thwart an undesired individual gaining access to the resources. But, the weakest links in any organizations security defences are its employees. Hundred percent security can never be achieved by just trying to prevent attacks on a technical level and not bothering about the physical-social level. The ignorance of this vital social element provided the hackers an easy method for obtaining access to a private system.

In Cyber Security terms, Social Engineering is the name given to the category of attacks involving purposeful manipulation of an individual or group in an effort to gain information or affect certain behaviour usually via some form of deceit and concealment of its actual objective. The term Social Engineering has been given many definitions related to both physical and cyber aspects of that activity. Some noteworthy definitions given by other authors are:

"An outside hacker's use of psychological tricks on legitimate users of a computer system, in order to obtain information he needs to gain access to the system".

"The practice of deceiving someone, either in person, over the phone, or using a computer, with the express intent of breaching some level of security either personal or professional".

"Social Engineering is a non-technical kind of intrusion relying heavily on human interaction which often involves tricking other people into breaking normal security procedures".

Social engineering attempts can be broadly classified into two categories namely technology based deception and human based deception. In technology based approach, the user is tricked to believe that he is interacting with a real application or a system thereby divulging confidential information and allowing access to an organizations network. In Human based approach, attacks are carried out by taking advantage of predictable human responses to psychological triggers.

People using social engineering techniques to research and collect data for illicit purposes are called social engineers who might be script kiddies, malicious hackers, Cybercriminals, Disgruntled employees, Terrorists, Scam artists or general people. The secret behind a truly successful social engineer is that they gather information without raising any suspicion as to what they are doing.

Though each one of the social engineering attacks is unique, there exists a common pattern for any social engineering attack. This pattern usually

termed as a cycle is given below and it generally consist of four phases.

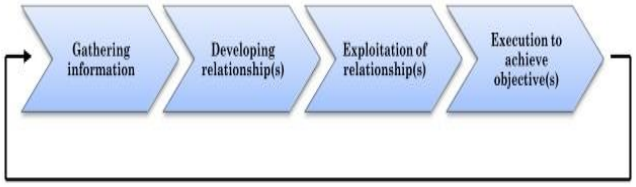


Fig. 1. Life Cycle of a Social Engineering Attack

The first phase is where the attacker gathers information about the target and the surrounding environment (also called footprinting). After enough information is gathered, the attacker now moves to the second phase, where he tries to develop a trust, building a rapport with the target (called manipulation phase). During the third phase, the attacker manipulates the trust gained in the previous phase and starts extracting sensitive data or operations. In the final stage, the attacker makes a clear exit in such a way that no proof is left behind and nothing could lead a trace back to his real identity thus completing the cycle.

II. ORIGINS OF SOCIAL ENGINEERING

Social engineering attacks are not new to this age. Long before the internet, electronic age or even the industrial revolution, they existed. Maybe it started with the first ever lie told. Some form of deception is always being employed in warfare from the start of warfare itself. The best example is that of the Trojan horse employed during the Greek-Trojan war, which ultimately resulted in a sound Greek victory over the Trojans. Now as we entered into the cyber era, social engineering has evolved by leaps and bounds and poses a formidable threat to the individuals and organizations.

III. COMMON WAYS OF ATTACK

Social engineering attacks happen in many different ways and can be performed anywhere when human interaction is involved. They can be both technical as well as non-technical in nature. The variation and extent of social engineering attacks are only limited by the creativity of the hacker. This section provides information about some of the most common forms of social engineering assaults.

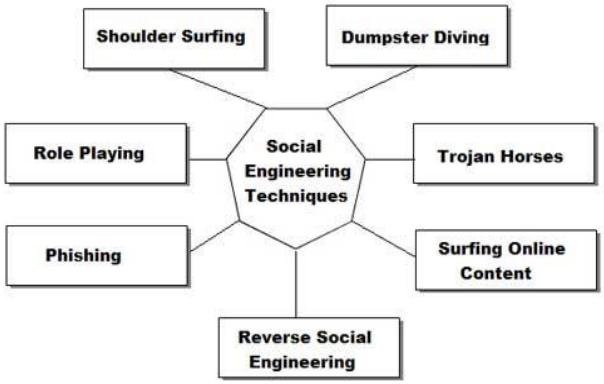


Fig. 2. Some Common Social Engineering Attacks

A. PHISHING

Phishing has been the most prolific form of social engineering and the number of victims is always on the rise. Phishing involves creating websites and emails that are carefully designed to look just like the legitimate ones. These trick the user into disclosing their personal information. While Email phishing remains the most widely used phishing attack, it can also be carried out by phone calls, text messages or even through social media.

One of the recent phishing scams involved receiving phishing emails by users who installed cracked APK files from Google Play Books, which contained malware. Once these infected books are downloaded, they contained instructions which redirect the users to a site, which attempts to load suspicious EXE files and unrelated malware APKs onto the phone. The most recent notable phishing attack was reported by Pivotal Company, where the attacker apparently netted a considerable amount of personal employee data. This attack involved an email purportedly by Pivotal CEO sent to the employees requesting payroll information. One of the recipients thinking it to be legitimate has forwarded it to all the other employees. After discovering that they have been duped, the company is taking countermeasures to safeguard the employee personal data.

A New kind of phishing tactic called as Chat-in-the-middle involves addition of a bogus live chat support window, where the user is enticed to enter their usernames and passwords in that chat session initiated by the fraudster.

The Anti-Phishing Working group reported a 250 percent increase in the number of phishing websites it detected during October 2015 to March 2016. With the attacker getting bolder every day, the attacks are also becoming more sophisticated and effectively targeted.

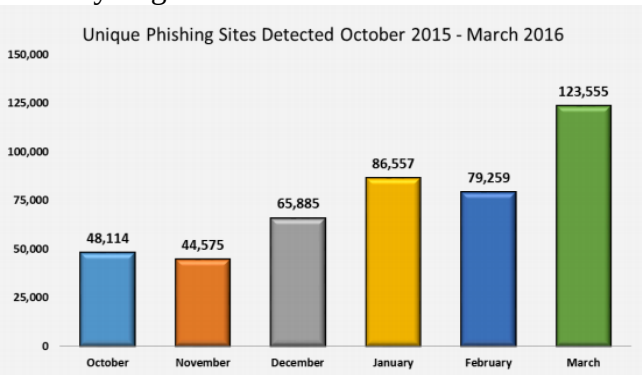


Fig. 2. Unique Phishing sites detected - Source: APWG Phishing Activity Trends Report

As internet continues to become an integral part of human life, the cyber threats are also on the rise. As the human tendency to trust their online systems and interact via social websites has grown over the years, phishing attackers also came up with new ways to exploit the human weaknesses. Two such notable ways are vishing and smishing.

Vishing (voice or VoIP phishing) is an electronic fraud scheme whose objective is to extract bank details or personal information from the victim and makes use of voice technology. It can be carried out by voice email, automated dialling, VoIP (voice over IP), landline or even a human on the other end of the phone. The victims are then lured into divulging their confidential information like credit card numbers, pin numbers, account numbers, passwords etc. These attacks commonly involve a scheme called as caller ID spoofing making the calls look like coming from a trusted source like a bank or law enforcement agencies. With the advent of VoIP, vishing attacks have become literally untraceable for the authorities. In April 2016, a call supposedly from fire department made employees of burger king, Minnesota break their windows so as to release the pressure buildup inside the store.

Smishing is a combination of SMS and phishing which uses SMS messages to defraud an individual. Bogus text messages masqueraded as threats or offers from legitimate sources like banks, stores etc lure the individuals to enter their personal data and ultimately become a victim.

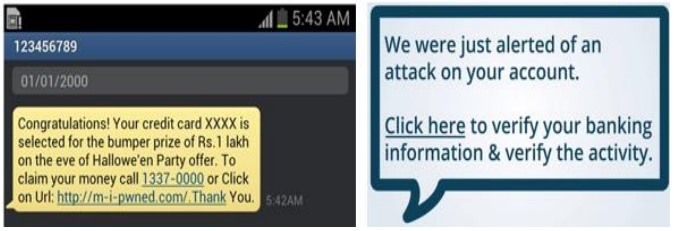


Fig. 2. Examples of Smishing SMS

With more and more people relying on their smart phones to access corporate data and networks, smishing has become the easier way to perform an attack for the criminals.

A new type of phishing, called spear-phishing or whaling is a more targeted approach which targets at employees or high-profile targets in a business. The attacker tricks the user into clicking a link or an attachment, which enables the attacker to create a backdoor to the targeted system. Now the attacker will be in a position to steal anything from the user ranging from corporate credentials, employee records, sensitive passwords and financial secrets. According to a survey by cloudmark, spear phishing is proving to be an expensive attack on the company. Some of the widely publicized spear phishing attacks are on JPMorgan Chase & Co., eBay, Target, Anthem, Sony and various departments within the U.S. government.

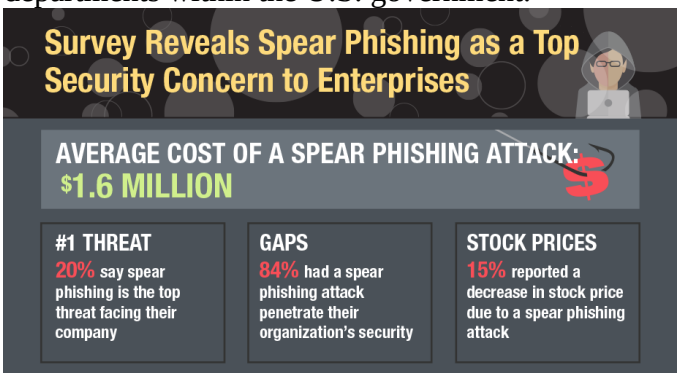


Fig. 4. Spear phishing survey - Source: Cloudmark

B. Baiting

Baiting is in a way similar to phishing, but makes uses of physical means rather than electronic means to carry out the attack. Baiting attack mainly relies on human curiosity or even greed of the victims. Generally, a piece of portable electronic storage like CD-ROM, USB stick, which is intentionally

left behind in the vicinity of the intended target (common areas like bathroom, parking place, lift, notice boards etc) with a tempting label or a file name on it. Because of the human tendency to find out what is on that disk, the victim inserts it into his computer, thereby unwittingly launching the malware present on the device. In most cases, in order to avoid suspicion, the attacker places other files such as music, game installation files so that the user doesn't suspect that he is being tricked. Once the malware is inserted, it provides access to the victims PC or the company internal network for the hacker.

C. Dumpster Diving

Dumpster diving is considered as a low-tech attack, but with serious implications. This term refers to examining waste and discarded products to find useful information. This attack is famous in 1980's as the security is lax and dumpster divers used to dive in the places like recycle bins, dumpsters, trashcans etc, where they get hold of the information for free. The information can be obtained from anything like company manuals, password files, diskettes, sensitive documents, credit card numbers, receipts, or financial reports. Later this information can be used by fraudsters for performing identity fraud. A famous example of dumpster diving is done in 1970's by a teenage hacker named Jerry Neal Schneider who collected a lot of information from Pacific Telephone and Telegraph (PT&T) company in Los Angeles. He posed as a freelance magazine writer, got a tour of the company and information about ordering procedures. He was able to steal about $1 million dollars worth of equipment and was only caught because of the tip from one of his disgruntled employees. Most of the corporate companies and Defence departments have identified the potential of this attack and designed policies that require secure disposal of garbage using shredding and even burning.

D. Shoulder Surfing

Shoulder Surfing means watching someone use their computer from over the shoulder. This technique allows the attackers to catch sensitive information and passwords by watching a user. This is a passive kind of technique which can be done physically or remotely (by cameras or by software).

E. Tailgating

Tailgating is another form of social engineering attack also known as piggybacking. Here, the attacker seeking entry into a restricted area can actually walk behind a person having legitimate access. Sometimes, in case of medium sized offices, the attackers try to strike up conversations with employees and later use this relationship to get past the front desk. The ultimate goal in this kind of attack is to get physical access to the site.

F. Quid Pro Quo

In Latin, Quid Pro Quo translates to 'this for that'. This attack promises the victim a benefit for exchange of information. The attackers pose as professionals and offer IT assistance to the users. Thinking that providing network credentials is required to solve the problem, users provide them to the attacker giving him all the access he needs. Quid Pro Quo tactics also extend beyond IT fixes. Surprisingly, real world examples have shown that employees revealing their network credentials for free gifts such as candy and pens.

G. Pretexting

It is the type of social engineering attack which is targeted and involves inventing a scenario to gather information from an unsuspecting user. The attacker researches about the target and collects enough information to use it for manipulation or impersonation. For this attack to be successful, a solid pretext is necessary. The key things- research, information gathering and planning results in building a solid pretext and a successful attack. One example of using pretexting for investigative purposes is by a group named Perverted Justice, who pose as young girls to lure paedophiles in and get them arrested for their perverted ways. They had their 600th successful conviction in April 2016.

H. Reverse Social Engineering

Reverse Social Engineering is a unique kind of attack as rather than the attacker going to the victim to gather information, here the victim unwitting goes to the attacker. This involves three steps- sabotage, advertising and assisting. The idea behind

this is that the attacker initially sabotages the network. Then, he advertises himself posing as someone from a legitimate organization such as a tech support service capable enough to solve the problem. When the victim sees the advertisement and thinking the attacker as the genuine consultant, welcomes him and allows him to work on the system or the network. With web based online services such as Facebook, Twitter & LinkedIn growing more and more every day, they are often becoming the target of the attackers to launch a wide automated reverse social engineering attacks. A lot of research has been going on trying to prevent such attacks on social networking sites.

I. Ransomware

Ransomware has emerged to become the most dangerous cyber threat for both organizations and consumers leading to annual losses of about hundreds of millions of dollars. It can be quoted to be the finest and the worst of all social engineering attacks. It involves installation of a malware which holds a user's computer hostage until a ransom fee is paid. Once compromised, the ransomware can either typically encrypt the key files or completely lock out a user from his computer. Generally, the ransom payments are always demanded in bitcoins –a form of digital currency, created and held electronically with no one centrally controlling it.

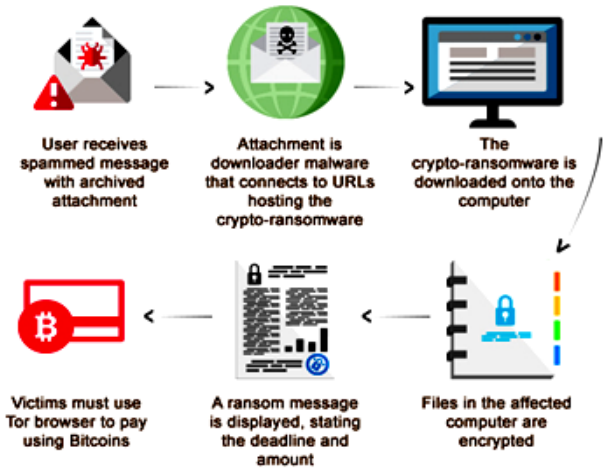


Fig. 5. Common scenario of Ransomware attack

Though there are many ways of distributing the malware or the Trojan horse, high percentage of distribution is through clicking on links attached to phishing emails.

According to the Symantec report, today the average ransom demanded from the victims has grown up to $679. No kind of organization be it a software company, law firm, manufacturing units, Government agencies, Police stations, Hospitals, Schools, Offices and even home users are immune to ransomware attacks.

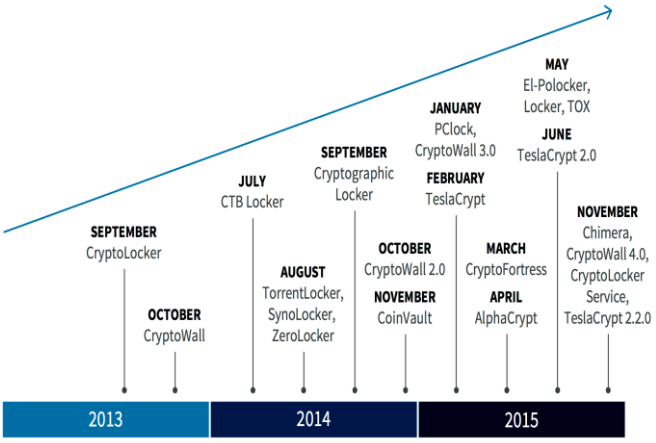


Fig. 6. Ransomware families, 2013-2015 (Source: Bromium)

There are two common forms of ransomware today namely Locker Ransomware (which denies access to a user's computer or a mobile device) and Crypto ransomware (prevents access to files or data). Symantec 2016 report highlights the shift towards Crypto ransomware in the last few years. One notable ransomware attack is Hollywood Presbyterian Hospital in Los Angeles paid more than $17,000 in bitcoin during February. As these attacks proved to be lucrative, they are becoming more and more sophisticated. The following statistics indicate that attackers are targeting developed, affluent nations.

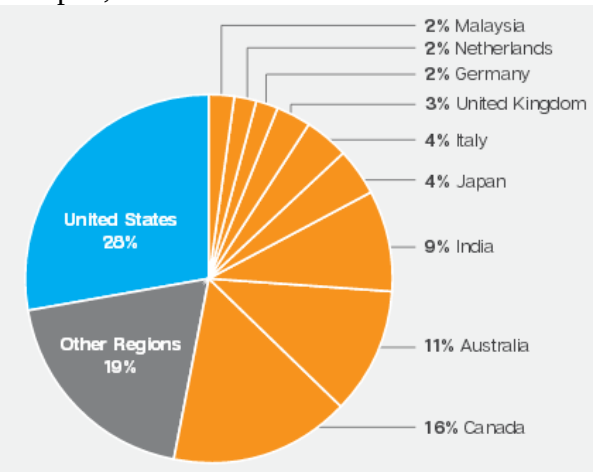


Fig. 7. Ransomware infections by region Jan15 – April 16 (Source: Symantec)

IV. DEFENSE AGAINST SOCIAL ENGINEERING ATTACKS

The above section provides an overview about the different kinds of techniques which are employed by social engineers. Some of them are based on technological means while others use various means of human manipulation. The best defense against any social engineering is UYCS (Use Your Common Sense) technique. Only when the people are educated about the kind of attacks being implemented and countermeasures to be used, they will not be in a position to defend themselves. As per M. Perlman, Eight general suggestions can be made regarding safeguarding from a social engineering attack.

1. Never divulge any kind of confidential information and network credentials via email, phone or in person to any unknown or suspicious parties.
2. Don't click on any suspicious attachments received via emails, though they look like coming from your contacts. Check the email address to see if it is legitimate or not.
3. When clicking on links, check for misspellings, @ symbols and suspicious sub-domains
4. Whenever you find something being downloaded automatically after clicking a link or visiting a website, report such activity immediately to the concerned authority.
5. Network administrators should constantly look out for any private or confidential information being posted on the company websites mistakenly by the users and remove them.
6. Block access to external storage devices to safeguard against baiting. If an infected device is plugged into the network, the entire network can be hacked.
7. Implement Awareness, Training and Education security concept for all the employees in the organization. Create suitable security policies and implement them strictly.
8. Use a multi-factor authentication for employees such as keycards & passwords along with a bio-metric password to make it difficult for the hackers to break into the organization.

In addition to the above listed suggestions, there are some other additional mechanisms which could be helpful in keeping the attackers at bay. Simple things such as constant updating of software, Good Intrusion detection systems (IDS), Secure disposal of waste (documents, media, paperwork etc.), strong password policies, proper procedures for verification of identities of users to IT personnel and vice-versa could help a lot to safeguard against social engineers.

V. CONCLUSION

The threat posed by social engineering is very real and cannot be ignored. Even though these scams have been going on and on for years, still people are becoming victims to these attacks. The main drawback is the lack of basic cybersecurity training available to the individuals. Not even the best designed network defence could stop a social engineering attack from happening because of the involvement of the human element. Though software has been developed to combat some techniques, a profound overall defence could only be achieved by a corporate wide culture of security awareness, which helps employees to routinely identify and repel social-engineering attacks. Also, training for the users on some of the tools used by hackers like the The Social-Engineer Toolkit (SET) could be imparted.

REFERENCES

[1] Michael Alexander, "Methods for Understanding and Reducing Social Engineering Attacks", SANS Institute InfoSec Reading Room
[2] Thomas R. Peltier "Social Engineering: Concepts and Solutions". Retrieved on August 8th 2016 from http://www.infosectoday.com/Norwich/GI532/Social_Engineering.htm
[3] Mandy Page, "Social Engineering: Information Bandits", SANS Institute Retrieved on August 11th 2016 from https://www.giac.org/paper/gsec/4202/social-engineering-information-bandits/106723
[4] Brandon Atkins et al. "A Study of Social Engineering in Online Frauds", Open Journal of Social Sciences, Vol.1, No.3, 23-32, 2013
[5] Katharina Krombholz et al., "Advanced Social Engineering Attacks", Journal of Information Security and Applications, 2014
[6] Nate Lord, Social Engineering Attacks: Common Techniques & How to Prevent an Attack, June 2016. Retrieved on July 20th 2016 from https://digitalguardian.com/blog/social-engineering-attacks-common-techniques-how-prevent-attack
[7] Manske, K. (2000). "An introduction to social engineering. Information Systems Security"
[8] Jared Kee, "Social Engineering: Manipulating the Source", SANS Institute Infosec Reading Room, 2008
[9] Ezer Osei Yeboah-Boateng et al., "Phishing, SMiShing & Vishing: An Assessment of Threats against Mobile Devices", Journal of Emerging

Trends in Computing and Information Sciences, Vol 5, No 4 April 2014

[10] Raj Samani, Charles McFarland, "Hacking the Human Operating System", McAfee Labs and Intel Security, 2015

[11] Whitwam, R. (2015). Google Play Books Is Crawling With Fake 'Guides' That Promise Cracked Android APKs, Provide Only Malware And Phishing Scams. Android Police. Retrieved August 11, 2016 from http://www.androidpolice.com/2015/03/03/google-playbooks-is-crawling-with-fake-guides-that-promise-cracked-android-apks-provide-onlymalware-and-phishing-scams/

[12] Barb Darrow (2016). Another Company Succumbs to Phishing Attack. Retrieved August 13, 2016 from http://fortune.com/2016/03/25/pivotal-phishing-attack/.

[13] Jonathan Crowe (2016). Phishing by the Numbers: Must-Know Phishing Statistics 2016. Retrieved August 20, 2016 from https://blog.barkly.com/phishing-statistics-2016 (2007) The IEEE website. [Online]. Available: http://www.ieee.org/

[14] Danesh Irani et al., Reverse Social Engineering Attacks in Online Social Networks, DIMVA 2011, LNCS 6739, pp 55-74, Springer 2011

[15] An ISTR Special Report: Ransomware and Business 2016 by Symantec, Retrieved August 5th 2016 from http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/ISTR2016_Ransomware_and_Businesses.pdf

[16] David Gragg, "A Multi - Level Defence Against Social Engineering", SANS Institute Infosec Reading Room, 2003. Retrieved Aug 20 2016 https://www.sans.org/reading-room/whitepapers/engineering/multi-level-defense-social-engineering-920

[17] Protect Against Social Engineering - Security-Aware Culture Helps Neutralize Social-Engineering Threats by Cisco. Retrieved Aug 2nd 2016 from http://www.cisco.com/c/en/us/about/security-center/protect-against-social-engineering.html

[18] M. Alexander, "Methods for Understanding and Reducing Social Engineering Attacks", SANS Institute Infosec Reading Room, 2016. Retrieved Aug 5th 2016 from https://www.sans.org/reading-room/whitepapers/engineering/methods-understanding-reducing-social-engineering-attacks-36972

[19] Menachem Perlman, "8 Tips to Prevent Social Engineering Attacks", Cyber Security Blog, 2014. Retrieved Aug 4th 2016 from http://lightcyber.com/8-tips-to-prevent-social-engineering-attacks/

A Survey on Big Data Analysis for High Velocity Data

K.S.Vijayalakshmi
Asst. Professor: CSE Department
VRSEC
Vijayawada, India
vijaya@vrsiddhartha.ac.in

Dr.K.V.Sambasiva Rao
Dean: CSE Department
NRI Institute of Technology, Agiripalli, India.

Vamsi Nadella
Grad Student: University of Georgia
Athens, Atlanta, USA
nadellavamsi94@live.com

Dr.E.V.Prasad
Professor: CSE Department
Director: LBRCE, Mylavaram, India.

Dr. V. Srikanth
Director: Citi Bank
London, UK.

Abstract— **Big data a word getting greater attention during the time not only in computer science community but also in all the entities with a minimum computer knowledge. The technology which is expected to have exploding quantity of applications and opportunities in the near future is defined with a non-slandered definition, but the one with N number of V's getting more acceptability. Though the Number of V's(N) changes from one to another. Initial era used a definition with 3 V's and later with 5 V's and now people are defining big data with 7 V's (Volume, Velocity, Variety, Variability, Veracity, Visualization and Value) [1]. So to define big data in this paper we are concerned with a definition that handles data with the speed it gets updated like, Big data can be defined as data that can't be handled by a traditional Database Management System and can be further defined as a technique which classifies data into Batches or Streams, where Batches are pre- recorded data which are divided into sub parts basing on a property of data and streams are like the data that gets updated continuously basing on time or any event which may be physical or logical that occurs in real-time. This paper illustrates various types' data that ranges from batch to high velocity and various big data frame works to deal with that data.**

Keywords- framework, streams, high velocity.

I. INTRODUCTION

The pace at which we generate data is growing gradually, thus creating even larger streams of endlessly evolving data. Online news, micro blogs, search queries, stock trades are just a few examples of these continuous streams of user activities. The value of these streams relies in their freshness and relatedness to ongoing events.

An *event* is the notice that an incident of interest has occurred. Event streams are constant flow of events generated from one or several producers. Substantial data streams that were once unclear and distinct are being aggregated and made easily accessible. Modern applications consuming these streams require extracting behavior patterns that can be obtained by aggregating and mining statically and dynamically vast event histories [2].

Processing streams can be considered as programming paradigm. This paradigm involves implementation of efficient algorithms, fault-tolerance mechanisms and methods to keep things happening. Timely analysis of such streams can be profitable in fields like finance, helpful in fields like agriculture and can even save lives in health care.

The main problem with streams is that, the event streams occur at higher speeds and in some cases with large volume of data that cannot be stored on a disk. In order to overcome this scenario, the algorithms should process them in single time under very strict constraints of space and time. Streaming algorithms use probabilistic data structures and give fast, approximated answers. However, sequential online algorithms are limited by the memory and bandwidth of a single machine. Achieving results faster and scaling to larger event streams requires parallel and distributed computing.

From the above situations the processing methods and infrastructures are to be set up that can convert a high volume raw stream into useful less volume data by either reducing the cardinality or arity. Numerous solutions can be used, for example stream database systems were very trendy research areas few years back. Commercial implementations of stream database like Streambase or Truviso allows users to use declarative language which is a derivative from SQL or continuous event streams.

Later ages came up with systems relevant to todays what so called Big Data Velocity.

II. FRAMEWORKS THAT SUPPORT LOW-LATENCY ANALYSIS

A. *Apache Drill:*

Apache Drill [3] is inspired by Google's Dremel, which is low latency distributed engine for handling large scale datasets, including structured and semi-structured/nested data. It is also useful for short and interactive queries on large data sets. Drill doesn't require a central meta-data repository and can query on nested data in formats like JSON and performing dynamic schema discovery. Drill can maximize data locality without moving the data among network or nodes. Relational data in Drill is treated as a special or simplified case of complex or multi structured data. When we view Drill in architectural point of view it can provide a flexible hierarchical columnar data model that can represent complex, highly dynamic and evolving data models.

Drill is designed from scrap for high performance and low latency. It doesn't use a traditional or general purpose execution engine like MapReduce or Tez. For these reasons it is flexible and performant. Drill supports a columnar and vectorized execution engine which results in high memory and CPU efficiency.

Drill consists of a service called Drill bit at the core of the architecture. This service can be installed in every node in a cluster to form distributed cluster environment. It provides an extensible architecture for all layers including storage plugin, query, query optimization and execution and client API layers. We can tailor any layer according to the particular needs of organization.

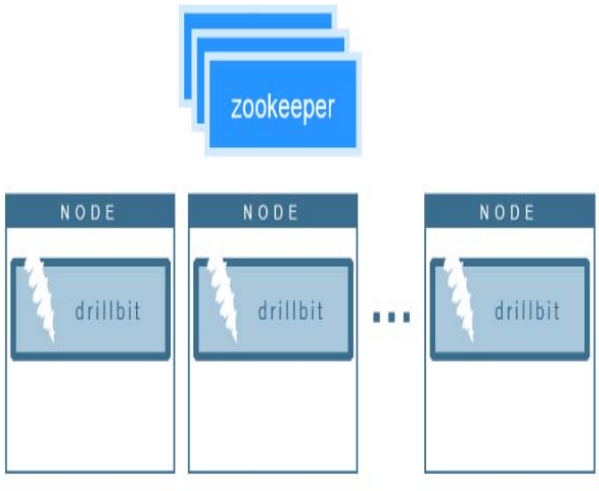


Figure 1. A Sample of overall Architecture for Apache Drill

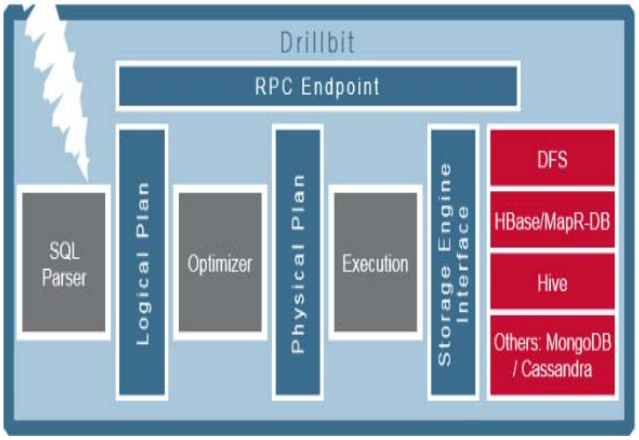


Fig 2. A Diagrammatic representation of a DRILL Node

Scalability of Drill is so huge that it can be implemented on a single personal computer and can scale up to a thousand node cluster. It can be used over Hadoop layer and it also includes a distributed execution environment. As mentioned above Drill does not have a centralized meta-data repository. It gathers its meta-data from storage plugins that keep up a correspondence with data sources. Storage plugins provide a variety spectrum of metadata like full meta-data, partial meta-data and decentralized meta-data which means Drill does not depend on a single Hive for meta-data. When a query is made the data from multiple repositories is collected and combined with information from HBase tables or within in a file in a distributed system.

III. FRAMEWORKS THAT SUPPORT STREAMS

A. *IBM Streams:*

InfoSphere® Streams [4] consists of a Programming Language, an API, and an Integrated Development Environment (IDE) for applications, and a runtime system that can run the applications on a single or distributed set of resources.

This framework was intended to address the following data processing platform objectives:

- Parallel and high performance streams processing software platform that can scale over a range of hardware environments.
- Automated exploitation of streams processing applications on configured hardware.
- Incremental exploitation without restarting to expand streams processing applications.
- Protected and auditable run time environment.

InfoSphere Streams provides a runtime platform, programming model, and tools for applications that are essential to process constant data streams. The need for such applications arises in environments where information from one to many data streams can be used to aware humans or other systems, or to populate knowledge bases for later queries. This architecture represents a significant change in computing system organization and capabilities.

InfoSphere Streams offers the IBM® Streams Processing Language (SPL) interface for users to operate on data streams. SPL provides a language and runtime framework to support streams processing applications. Users can create applications without the knowledge of lower-level stream-specific operations. SPL provides various operators, the capability to bring in data from outside InfoSphere Streams and communicate results outside the system, and an ability to expand the original system with user-defined operators. Several SPL built-in operators offer powerful relational functions such as Join and Aggregate.

Starting with InfoSphere Streams Version 4.1, users can also develop streams processing applications in other supported languages, such as Java™ or Scala. The Java Application API (Topology Toolkit) supports creating streaming applications for InfoSphere Streams in these programming languages.

Deploying streams processing applications results in the creation of a dataflow graph which runs across the distributed run time environment. As new workloads are submitted, InfoSphere Streams determines where to best set out the operators to meet up the resource necessities of both recently submitted and previously running specifications. InfoSphere Streams constantly monitors the state and consumption of its computing resources. When streams processing applications are going on, they can be monitored dynamically across a distributed group of resources by using the Streams Console, Streams Studio, and Stream tool commands.

Results from the running applications can be made available to applications that are running external to InfoSphere Streams by using Sink operators or edge adapters. For example, an application might use a TCP Sink operator to throw its results to an external application that visualizes the results on a map. Alternatively, it might alert an administrator to unusual or interesting events. InfoSphere Streams also provides many edge adapters that can connect to external data sources for consuming or storing data.

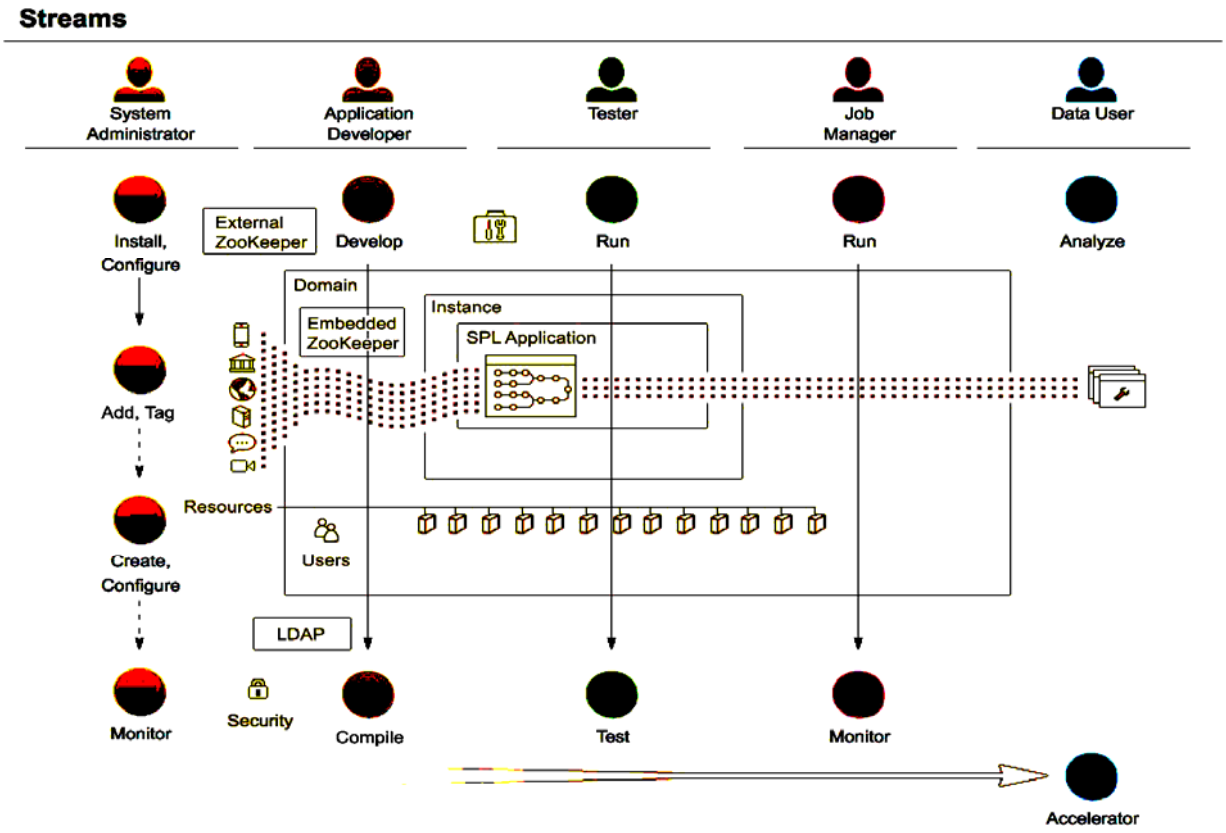


Figure 3. IBM Streams Architecture

B. Apache Storm:

Storm a rock star well known to people with big data knowledge. This framework is real-time distributed framework that provides fault-tolerance and guarantees data processing.

Storm is project developed from ground using Clojure and Java and usable with any other programming language. Storm can be used for different use cases:

- Stream Processing: Storm can be used to analyze new data and update the database in real time.
- Continuous Computation: Storm has great fault-tolerance mechanisms that provides continues query and results to clients in real time.
- Distributed RPC: Strom uses Di-Acyclic Graphs(DAG) for parallel processing and has a granularity of configuring deciding the number of threads in client systems.

Storm has the capabilities of both stream processing and batch processing. This was made possible by Lambda Architecture developed by main developer of Storm Nathan Marz.

Lambda Architecture:

This architecture [5] has the capabilities of mixing batch-processing and real-time data processing. This approach is divided into three layers:

1. The Batch Layer
2. The Serving Layer
3. The Speed Layer

The mechanism of Lambda Architecture is as follows:

When any new data is arrived, it will be sent to both the batch layer and the speed layer. In the batch layer, new data is added to the master data set. The master dataset is a set of files in HDFS and posses the raw information that is not derived from any other information. It is an immutable append-only set of data.

The batch layer pre computes query functions from scratch continuously, in a "while (true)" loop. The results of the batch layer are called "batch views".

The serving layer indexes the batch views formed by the batch layer. Basically, the serving layer is a scalable database that *swaps* in new batch views as they are made available. Due to the latency of the batch layer, the results available from the serving layer are always out of date by a few hours.

The speed layer compensates for the high latency of upgrades to the serving layer. This layer utilizes Storm to process data that have not been considering in the last batch of the batch layer. This layer generates the real-time views that are always that are dependably a latest mode and stores them in databases that are useful for both read and write. The speed layer is more intricate than the batch layer but that complication is remunerated by the fact that the real time views can be continuously disposed of as data makes its way through the batch and serving layers.

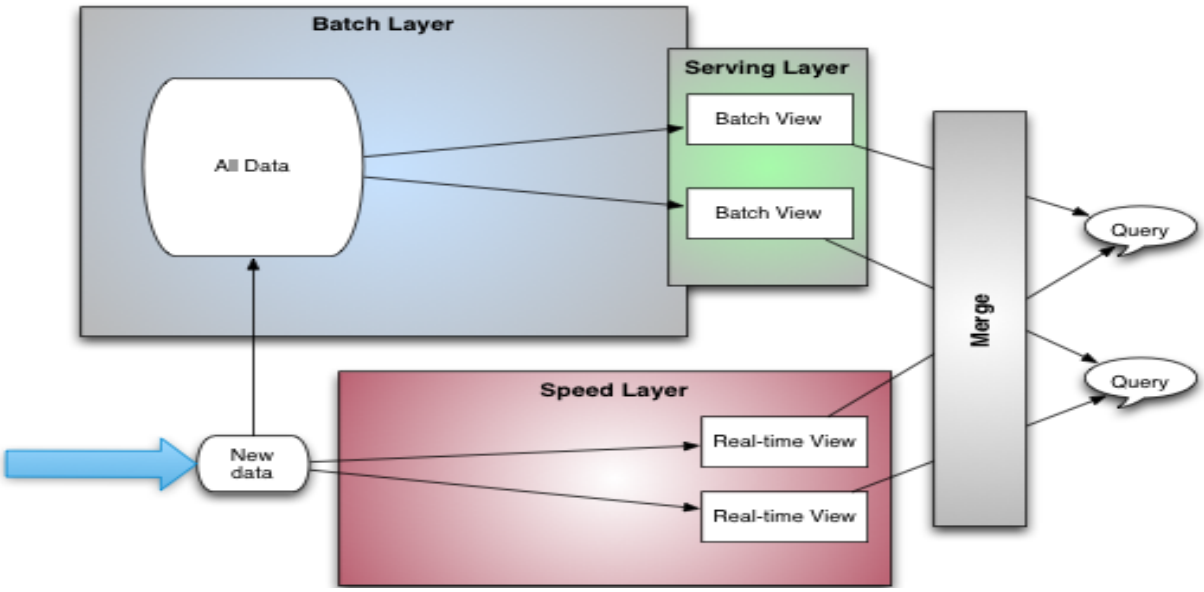


Figure 4. Lambda Architecture.

C. Apache Flink:

Flink [6] is framework that can both process streams and batch processing. Flink is specialized for stream processing and considers batch processing as a specialized case of stream processing. The major advantages of Flink are

- High Performance and Low Latency
- Support for Event Time and Out-of-Order Events
- Exactly-once Semantics for Stateful Computations
- Continuous Streaming Model with Backpressure
- Fault-tolerance using Lightweight Distributed Snapshots

Flink comes up with several Libraries for Machine Learning and Graph Processing. Flink consists of Query Processer to optimize queries similar that present in RDBMS but much more in a faster way.

It has its own way of memory management apart from JVM garbage collector and it comes up with mechanism of handling the clusters reducing the user tuning. For extra speed, Flink permits iterative processing to occur on the same nodes as opposed to having the cluster run every iteration autonomously.

Flink even supports out-of-order data streams which are caused when using data streams caused by distributed data producers or the data that travels in different paths and reaches the receiver in different order from the initial order.

Even before the first version of Flink being released it is more efficient than Spark in case of streaming. Spark got ruled out of streaming because it is not a pure stream oriented analysis framework. Spark just considers streams as batches with smaller sizes and smaller time periods which are other-wise called micro-batches.

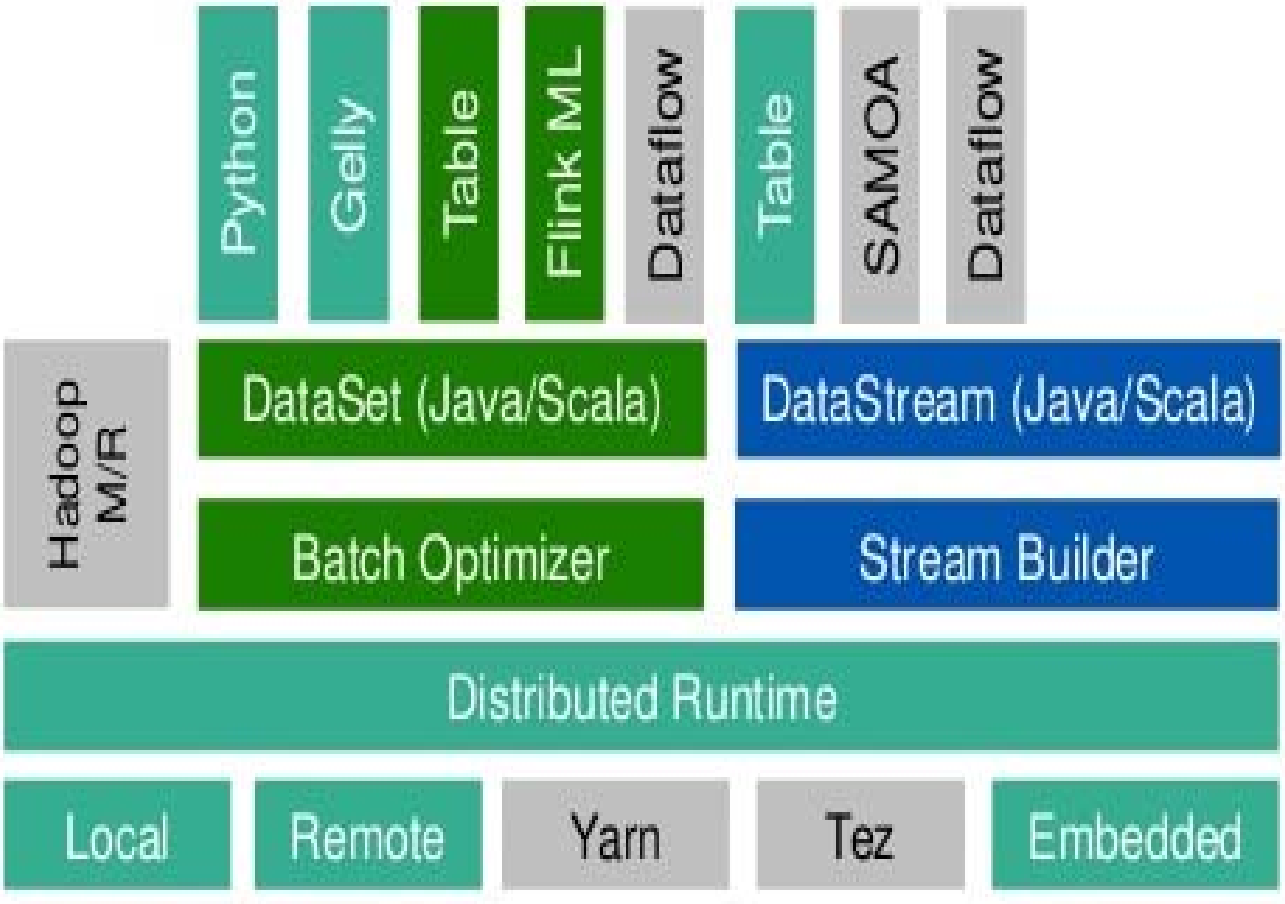


Figure 5: A Sample Stack of Apache Flink

CONCLUSION

This survey majorly concentrated on methods, techniques and frameworks for big data analysis of high frequency data. We tried to discuss several mechanisms for stream processing and a framework for low-latency analysis. Even though Apache Spark being said as both batch and stream processing framework, but it is not generically a stream processing framework because it considers streams as micro batches which may be fine with applications having relatively low velocity data but it doesn't work with high frequency applications like stock markets. We discussed about the above statement earlier in the paper.

Of the frameworks discussed so far for analysis of streams, every framework got different techniques for handling the data efficiently. IBM streams come up with a good support and interoperability with several open source tools and majority of the IBM data analytics products. An open source version of IBM streams will be available with methods of successful registration. Apache Flink being a next best and newly developed framework, has the majority capabilities of being the best solution with several advantages like huge library base, ability to process out-of-order data and fault-tolerance etc. Apache Storm which is already widely known for fault tolerant data framework. The maintenance through DAG technique has profound capabilities of processing streams and batches through Lambda Architecture which provides opportunities for analysis on great range of data gamut. This mechanism is further helped in making Apache Storm guarantee for data processing. This Lambda Architecture is being implemented by existing frameworks and upcoming techniques. From the so far discussion, let us conclude that Apache Storm has enough capabilities to be a leader in big data analytics.

REFERENCES

[1]. http://dataconomy.com/seven-vs-big-data/
[2].Genoveva VargasSolar, Javier A. EspinosaOviedo, José Luis ZechinelliMartini , Big Continuous Data: Dealing with Velocity by Composing Event Streams. Book Chapter in Springer Link. http://link.springer.com/chapter/10.1007/978-3-319-27763-9_1
[3]. https://www.mapr.com/blog/apache-drill-architecture-ultimate-guide
[4].http://www.ibm.com/support/knowledgecenter/SSCRJU_4.1.1/com.ibm.streams.welcome.doc/doc/ibminfospherestreams-introduction-overview.html
[5]. Thibaud Chardonnens, Big Data Analytics on High-velocity Stream, Master Thesis, Department of Informatics, University of Fribourg (Switzerland), 2013.
[6]. https://disqus.com/home/forum/stratosphere-eu/

AUTHORS PROFILE

K.S.Vijayalakshmi, is an Assistant Professor at the Department of Computer Science and Engineering, V. R. Siddhartha Engineering College, Vijayawada, India. Her research interest is in the area of Data mining, Machine Learning and Data Analytics.

Vamsi Nadella, is a Graduate Student at the Department of Computer Science and Engineering, University of Georgia, Athens, Atlanta, USA. His research interest are in the area of Data Analytics and Machine Learning.

Dr.K.V.Sambasiva Rao, is a Professor and Dean at the Department of Computer Science and Engineering, NRI Institute of Technology, Agiripalli, India. His research interest is in the area of Decision Support System.

Dr. E.V.Prasad is a Professor of Computer Science and Engineering and Director at LBR College of Engineering, Mylavaram, India. His research is in the area of Parallel Processing, Data Mining etc.

Dr. V. Srikanth, is a Director at Citi Bank, London, UK. His research interest in the area of Machine Learning and Evolutionary Computing.

Comparative Study Review on Lung Cancer Detection Using Optimization and Clustering Approach

Divya Chauhan
M. Tech.
Shoolini University
Solan, India(H.P.)
divya.chauhan93@gmail.com

Varun Jaiswal
Assistant Professor
Shoolini University
Solan, India (H.P.)
computationalvarun@gmail.com

Abstract- Among different infections malignancy has ended up significant danger in India. According to Indian populace because of disease the death rate was high. Malignancy is the second most illness in charge of death in India. In perspective of these certainties, this paper will surveys two systems i.e. Advancement and Clustering Algorithm with their cons andpros, that are exceptionally useful in early analyze of lung tumor. Notwithstanding above, at last this paper will close which procedure is ideal and must be received for better exactness of malignancy avoidance framework.

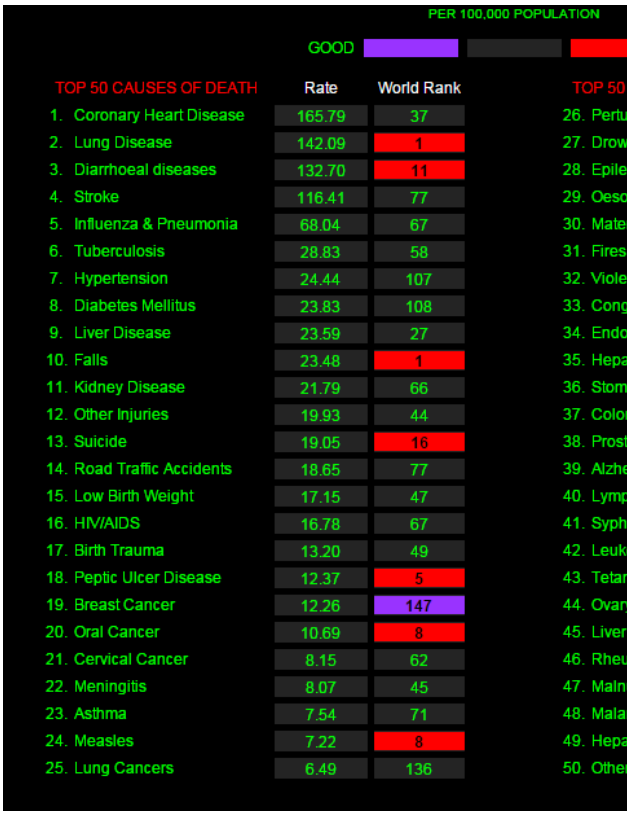
Keywords- Lung Cancer,Optimization, Clustering Algorithm, CT images

I. INTRODUCTION

The high pervasiveness of lung disease prompts its initial aversion. The presentation of PC innovation parcels in expanding the death rate of the lung malignancy patients because of its discovery at early stages. Figure out if a pneumonic knob is a being tumor or not in the early stages is vital. Be that as it may, determination of the nearness of tumors in little knobs is extremely troublesome. With the fast headway of the innovation, the collaboration between material science, designing and registering science has turned out to be nearer than any time in recent memory. A greater number of individuals kick the bucket on account of lung growth than whatever other sorts of malignancy, for example, Breast, colon, and prostate diseases as appeared in Figure.2. Human machine frameworks for picture based conclusion need to exploit both human and machine abilities, making a framework, which overall will be more noteworthy than the whole of its parts (Katherine et.al, 2003)In India by far most cases (90%) of lung malignancy is because of presentation to tobacco smoke. Around 10 % of malignancy happens in that individuals who never smoked. These cases are regularly brought about because of the hereditary impacts. Lund tumor is the most widely recognized reason for death in India and was in charge of 1.56 million passingevery year, according to overview in 2012 and in 1991 around 60, 9000 individuals was affected by lung growth.

Fig. 1 CTimage showing Lung cancer

Lung growth arranging is an evaluation of the level of spread of the tumor from its unique source. It is one of the components influencing the visualization and potential treatment of lung growth (Hornet.al, 2012). Underneath outline demonstrates the reasons of death in India. From chart it is plainly seen that Lung disease is at second generally put. Late studies demonstrates that individuals living at higher elevations has okay rate of lung malignancy in smoker and in addition in non-smoker, which demonstrates that oxygen may advance the lung tumor issue, This is distributed online in PeerJ. Additionally as indicated by Oncology Nurse Advisor, with the height of the elevations, lung malignancy rate tumbled to 7.23 cases in 1000 individuals.



PER 100,000 POPULATION

GOOD

TOP 50 CAUSES OF DEATH	Rate	World Rank	TOP 50
1. Coronary Heart Disease	165.79	37	26. Pertu
2. Lung Disease	142.09	1	27. Drow
3. Diarrhoeal diseases	132.70	11	28. Epile
4. Stroke	116.41	77	29. Oesc
5. Influenza & Pneumonia	68.04	67	30. Mate
6. Tuberculosis	28.83	58	31. Fires
7. Hypertension	24.44	107	32. Viole
8. Diabetes Mellitus	23.83	108	33. Cong
9. Liver Disease	23.59	27	34. Endo
10. Falls	23.48	1	35. Hepa
11. Kidney Disease	21.79	66	36. Stom
12. Other Injuries	19.93	44	37. Colo
13. Suicide	19.05	16	38. Prost
14. Road Traffic Accidents	18.65	77	39. Alzhe
15. Low Birth Weight	17.15	47	40. Lymp
16. HIV/AIDS	16.78	67	41. Syph
17. Birth Trauma	13.20	49	42. Leuk
18. Peptic Ulcer Disease	12.37	5	43. Tetar
19. Breast Cancer	12.26	147	44. Ovar
20. Oral Cancer	10.69	8	45. Liver
21. Cervical Cancer	8.15	62	46. Rheu
22. Meningitis	8.07	45	47. Maln
23. Asthma	7.54	71	48. Mala
24. Measles	7.22	8	49. Hepa
25. Lung Cancers	6.49	136	50. Othe

Figure. 2 Causes of Death in India
(www.lungindia.com)

There are numerous systems to analyze lung tumor for example, mid-section radiograph(x-ray), processed tomography (CT), attractive reverberation imaging (MRI output) and sputum cytology (Wang, 2006). Be that as it may, a large portion of the procedures are extremely costly. In this way there is incredible need of an innovation in which the death rate is exceptionally high.A number of therapeutic scientists used the investigation of CT pictures for early recognition of lung growth (Shiela, 2010), latest examination transfer on quantitative data, for example, the size, shape and the proportion of the influenced cells (Kim, 2005).

Thus, most analysts are attempting to build up an analyze framework on the premise of CT pictures. There are numerous calculations which have been proposed in medicinal imaging. Be that as it may, in this survey paper we will audit the most encouraging techniques like Optimizing and Clustering Algorithm (CM) with their preferences too inconveniences and their approach to analyze lung growth in people.

The indication of this paper is sorted out as takes after. In Section 2, Literature overview is introduced. In Section 3, streamlining calculation is de-scribed. In Section 4, bunching calculation is introduced lastly in Section 5, the conclusion and future work are given.

II. RELATED WORK

Fuzzy k-c-implies bunching calculation utilized for therapeutic picture division which was presented in (Ajala, 2012). Here fuzzy c-means is a strategy for grouping calculation which permits one bit of information has a place with two or more bunches and k-means is a basic bunching technique in which we utilize low computational multifaceted nature when contrasted with fuzzy c-implies. At the point when both Clustering techniques were consolidated to create an additional time proficient division calculation called as fuzzy k-c-implies bunching calculation. They offered that limit which is the most rudimentary system for restorative picture division, in which this calculation partitions pixels in various classes relying on their dark level. It is likewise said that it approaches division of scalar pictures by framing a double segment of the force estimations of a picture and in conclusion decides a power esteem. This power worth is termed as edge, which isolates the craved classes. Classifier procedures which were utilized for example acknowledgment, segments a component space got from the picture utilizing information with known marks. An element space is an arrangement of N*M lattice where N identifies with the quantity of perceptions and M identifies with the quantity of traits. Classifiers are known as administered techniques since they require preparing information which are physically divided and afterward utilized it for naturally portioning new information.

An examination between two strategies was made in (Christian, 2012). These techniques are tenet based strategy and Bayesian elegance technique for the extraction of cell area from foundation and flotsam and jetsam cell locale, and after experimentation the Bayesian style technique was discovered applicant this capable for arrangement of sputum cell district from foundation area. Yet, they didn't expel the core district from cytoplasm area with this procedure.

In this (Fatma, 2012) two more division strategies were utilized which were Hopfield Neural Network (HNN), and Fuzzy C-Mean (FCM) grouping calculation. In this they found that the HNN gives upgraded, precise and dependable division results than FCM grouping in all cases. The HNN likewise separates the cores and cytoplasm locales while FCM fizzled in the recognition of the cores. FCM just distinguished a part of the core not the entire core in a specific cell. Likewise FCM was not discovered

unpretentious to force varieties on the grounds that the division blunder at meeting was discovered bigger with FCM in contrast with HNN. As per the most extreme most recent assessments of the measurements which are given by world wellbeing association shows that there happened around 7.6 million passing's worldwide every year in view of this kind of tumor. In addition, they likewise found that mortality from tumor are evaluated to rise consistently, and will draw close to 17 million passing's worldwide in 2030. Thus, better techniques are required to extricate the core locale for early identification. A magazine in (IEEE, Pulse) gave us the learning about current patterns in therapeutic picture examination.

In (Mokhled, 2012) first pictures which were enhanced through Gabor channel. It has given preferred results over other upgrade procedures. They just chipped away at hued picture upgrade and not remove the core district and even not the cell locale. In Features Extraction stage they secure the general elements of the improved and sectioned picture which later they utilized as a part of binarization. A refined Charged Fluid Model (CFM) alongside enhanced Otsu's technique was utilized for the programmed division of MRI pictures in (Nagesj, 2012). This technique gave upgraded results than the outcome given by the methodologies utilized as a part of past tests.

In (Nikita, 2012), a calm edge recognition technique was utilized which depends on finding the picture inclination. This technique tells that force of the picture will be greatest where there is a detachment of two unique districts and subsequently an edge must exist there. On this premise they found the knobs in CT pictures.

In (Parsh, 2011), another variety level set calculation without re-introduction was utilized. They likewise utilized thresholding to decrease the clamor segment of the pictures.

In (Sajith, 2012) glandular cells were recognized by utilizing numerous shading spaces and two bunching calculations which were K-implies and Fuzzy C-implies.

In (Sonith, 2012) a review of whole process for handling computerized pictures for lung disease discovery is given in this paper. This paper likewise depicts all the fundamental strides required for the better execution beginning from the pre-preparing till the very end stage extraction of elements.

III. FIREFLY NETWORK

The Firefly algorithm is a freshly developed nature-inspired Meta heuristic algorithm that is an example of optimization algorithms. The Firefly algorithm is encouraged by the social presentation of fireflies. Fireflies may also be called lightning bugs. There are about 2000 firefly species in the globe. Most of the firefly species construct short and rhythmic flashes. The model of flashes is unique for a particular species. A firefly's twinkle mainly acts as a signal to attract mate partners and potential prey. Flashes also serve as a defensive warning instrument. The following three idealized rules are considered to explain the firefly algorithm (*K. Naidu, 2013):*

1) All fireflies are unisex so that one firefly will be involved to other fireflies despite of their sex.
2) Attractiveness is relative to their brightness; thus, for any two flashing fireflies, the less bright one will move in the direction of the brighter one. The attractiveness is relative to the brightness and they both reduce as their distance increases. If there is no brighter one than a particular firefly, it will move arbitrarily.

3) The clarity of a firefly is affected or unwavering by the landscape of the idea function. For a maximization problem, the brightness may be comparative to the objective function value. For the minimization problem, the brightness may be the give-and-take of the objective function value. The make believe code of the firefly algorithm was given by Yang*(M. H. Sulaiman, 2001).*

A. Attractiveness

The attractiveness of a firefly is determined by its light intensity. The attractiveness may be calculated by using the equation:

$$\beta(r) = \beta_o e^{-r^2}$$

B. Distance

The distance among any two firefly's k and l at Xk and Xl is the Cartesian distance as follows:

$$r_{ld} = \left\| x_k - x_l = \sqrt{\sum_{k=1}^{d} (x_{k,o} - x_{l,o}{}^2)} \right.$$

C. Movement

The movement of a firefly k that is attracted to another more attractive firefly l is determined by 0.

$$x_k = x_k + \beta_0 e^{-r^2}(X_l - X_k) + \propto (rand - \frac{1}{2})$$

Firefly algorithm is written as below:

Objective function $\boldsymbol{f(x), x = (x_1, \ldots . x_d{}^T)}$

Obtain original population of fireflies $x_i (i = 1,2,..n)$

$f(x_i)$ Is used to determine the light intensity $I_i$at x_i.

Define light absorption coefficient γ.

While (t<Max generation)

For i=1: n for n fireflies

For j=1:I for all n fireflies

If $I_j > I_i$, move firefly I towards j in d-dimension ;end

If

Attractiveness vary with distance r via $\exp(-\gamma r)$

Assess novel solution and inform light intensity

end for j

end for i

Rank the firefly and discover the present best

End while

Post process results and visualization

Advantages:

- Firefly can be accessed anywhere with a web browser and an internet connection. This makes it a potentially more convenient and portable tool than Kurzweil.
- The reading voices in Firefly are, as a whole, superior to those in the Mac version of Kurzweil.
- Firefly is comparable to Kurzweil (both Windows and Mac versions) in its ability to translate text. Unlike Mac Kurzweil, Firefly can also intelligibly read text in Spanish.

Disadvantages:

- Firefly is not a text editor, and it has limited annotation capabilities (Firefly only supports text highlighting).
- Firefly can only read .kes files. If you wish to upload a file of a different format, you must first create a .kes file using Kurzweil.
- As an online tool, Firefly is slightly slower than Kurzweil – it takes Firefly a moment to load text and move between pages.

IV. CLUSTERING APPROACH

Medical Data Mining is a promising range of computational insight connected to a naturally investigate patients records going for the revelation of new learning helpful for restorative basic leadership. Applying information mining methods to growth information is valuable to rank and connection tumor ascribes to the survival result. Further, exact result expectation can be greatly valuable for specialists and patients to gauge survivability, as well as help in basic leadership to decide the best course of treatment for a patient, in view of patient particular traits, as opposed to depending on individual encounters, accounts, or populace wide hazard appraisals.

Clustering is the subfield of data mining and it is the processof separating the data into identical regions based on the resemblance of objects; information that is logically related physically is stored together, in order to rise the efficiency in the database system and to minimize the number of disk access *(R.Duda, 2001).* The process of clustering is used to assign the q feature vectors into K clusters, for each k^{th}cluster$C^k$is its center. Fuzzy Clustering has been used in many fields like pattern recognition and Fuzzy identification. A variety of Fuzzy clustering methods have been suggested and most of them are based upon distance criteria (*Ramesh, 2011).* The most extensively used algorithm is the Fuzzy C-Mean algorithm (FCM), because it uses reciprocal distance to compute fuzzy weights. This algorithm has an input a pre-defined number of clusters, which is the *k* from its name. Here Means stands for an average location of all the members of particular cluster and the output is a partitioning of *k* cluster on a set of objects. The main objective of the FCM cluster is to minimize the total weighted mean square error (*Sun, 2004):*

Formula $J = (W^{qk}, C^{(k)}) = \sum \sum (W^q{}_k)^p \left\| x^{(q)} - c^{(k)} \right\|^2$

The FCM allows individually every feature vector to belong to multiple clusters using various fuzzy membership values. Then the final classification will be according to the maximum weight of the feature vector over all clusters. The detailed algorithm (Sun, 2004):

Input: Vectors of objects, each object represent *s* dimensions, where $v=\{v_1, v_2, \ldots\ldots, v_n\}$ in our case it will be an image pixels, each pixel has three dimensions RGB, K = number of clusters.

Output = A set of K clusters which minimize the sum of distance error.

1. Initialize random weight for each pixel, it uses fuzz weighting with positive weights $\{W^{qk}\}$between[0, 1].
2. Standardize the initial weights for each q^{th} feature vector over all K clusters via:

W_{qk} / W_{qr}

3. Standardize the weights over k = 1,…,K for each q to obtain W_{qk}, via: (*R.Duda, 2001)*

$W_{qk} = W_{qk} / \sum W_{qr}$, q = 1,…Q.

4. Compute new centroids $C^{(k)}$, k = 1,….,K via

$C^{(k)} = \sum W_{qk} x^{(q)}$, k=1,….K

5. Update the weights { W_{qk}, }
6. If there is any change in the input, repeat from step 3, or else terminate.
7. Assign each pixel to a cluster based on the maximum weight.

Advantages:

- Gives best result for overlapped data set.
- Data point is assigned membership to each cluster centre as a result of which data point may belong to more than one cluster centre.

Disadvantages:

- A prior specification of the number of clusters.
- With lower value of β we get the better result but at the expense of more number of iteration.
- Euclidean distance measures can unequally weight underlying factors.

V. CONCLUSION AND FUTURE SCOPE

The early detection of lung cancer is a challenging problem, due to the structure of the cancer cells, where most of the cells are overlapped with each other. This paper has presented two detection methods, optimization and clustering algorithm, for CT images to detect the lung cancer in its early stages. The manual analysis of the CT samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. This paper also presented methods with its advantages so that further work can be done according to the application.

RFERENCES

[1] Katherine P. Andriole, “Addressing thc Corning Radiology Crisis: The Society for Computer Applications in Radiology, Transforming the Radiological Intcrprctation Process (TRIP.) initsativc.”Position Paper from the SCAR TRIPTM Subcommittee of lhc SCAR Research and Development Committee, November 2003.
[2] www.lungindia.com
[3] Horn, L; Pao W; Johnson DH (2012). "Chapter 89". In Longo, DL; Kasper, DL; Jameson, JL; Fauci, AS; Hauser, SL; Loscalzo, J. Harrison's Principles of Internal Medicine (18th ed.). McGraw-Hill.
[4] http://www.worldlifeexpectancy.com/country-health-profile/india
[5] W. Wang and S. Wu, “A Study on Lung Cancer Detection by Image Processing”, proceeding of the IEEE conference on Communications, Circuits and Systems, pp. 371-374, 2006.
[6] A. Sheila and T. Ried “Interphase Cytogenetics of Sputum Cells for the Early Detection of Lung Carcinogenesis”, Journal of Cancer Prevention Research, vol. 3, no. 4, pp. 416-419, March, 2010.
[7] D. Kim, C. Chung and K. Barnard, "Relevance Feedback using Adaptive Clustering for Image Similarity Retrieval," Journal of Systems and Software, vol. 78, pp. 9-23, Oct. 2005.
[8] AjalaFunmilola A, Oke O.A, Adedeji T.O, Alade O.M, Oyo Adewusi E.A, “Fuzzy k-c-means Clustering Algorithm for Medical Image Segmentation”, Journal of Information Engineering and Applications, ISSN 2224-5782 (print) ISSN 2225-0506 (online), Vol 2, No.6, 2012
[9] Christian D., Naoufel W., Fatma T., Hussain, "Cell Extraction from Sputum Images for Early lung Cancer Detection", IEEE 978-1-4673-0784-0/12, 2012
[10] Fatma T., Naoufel W., Hussain, Rachid S., "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods", American Journal of Biomedical Engineering, 136-142 DOI: 0.5923/j.ajbe.20120203.08, 2012
[11] “Medical Image Analysis”, IEEE Pulse, 2154-2287/11/2011
[12] Mokhled S. AL-TARAWNEH, “Lung Cancer Detection Using Image Processing Techniques”, Leonardo Electronic Journal of Practices and Technologies, ISSN 1583-1078, Issue 20, January-June 2012
[13] Nagesh V., Srinivas Y., Suvarna Kumar G, Vamsee Krishna V, “An Improved Medical Image Segmentation Using Charged fluid Model”, International Journal of Engineering and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 2, pp.666-668, Mar-Apr 2012
[14] Nikita P., Sayani N., “A Novel Approach of Cancerous Cells Detection from Lungs CT Scan Images”, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN 2277 128X, Volume 2, Issue 8, August2012
[15] Parsh Chandra B., Md. Sipon M., Bikash Chandra S. and Mst. Tiasa K., “MRI Image Segmentation Using Level Set Method and Implement a Medical Diagnosis System”, Computer Science & Engineering: An International Journal (CSEIJ), Vol. 1, No. 5, December 2011
[16] SajithKecheril S, D Venkataraman, J Suganthi and K Sujathan, "Segmentation of Lung Glandular Cells using Multiple Color Spaces", International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.3, June 2012
[17] SonitSukhraj Singh, Anita Chaudhary “Lung Cancer Detection using Digital Image Processing”, IJREAS Volume 2, Issue 2 ISSN: 2249-3905, (February 2012)
[18] V.V. Thakare, P. Singhal,“ Neural network based CAD model for the design of rectangular patch antennas, ” JETR,vol. 2(7), 2010.
[19] R. Duda, P. Hart, ”Pattern Classification”, Wiley-Interscience 2nd edition, October 2001.
[20] S. Aravind, J. Ramesh, P. Vanathi and K. Gunavathi,” Roubust and Atomated lung Nodule Diagnosis from CT Images based on fuzzy Systems”, processing in International Conference on Process Automation, Control and Computing (PACC), pp. 1-6, Coimbatore, India, July, 2011.
[21] H. Sun, S. Wang and Q. Jiang, "Fuzzy C-Mean based Model Selection Algorithms for Determining the Number of Clusters," Pattern Recognition, vol. 37, pp.2027-2037, 2004.
[22] K. Naidua, H. Mokhli, A. H. A. Bakar, “Application of Firefly Algorithm (FA) based optimization in load frequency control for interconnected reheat thermal power system”, 2013 IEEE

Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), IEEE, 2013.

[23] M. H. Sulaiman, M. W. Mustafa, "Firefly Algorithm Technique for Solving Economic Dispatch Problem,"978-1-4673-0662-1/31.002012IEEE.

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Dr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Dr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2016
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/