3-3-2012

Proceedings of International Conference on Computer Science and Information Technology

Prof.Srikanta Patnaik Mentor
IRNet India, patnaik_srikanta@yahoo.co.in

Proceedings

of

International Conference on

COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

(CSIT 2012)

March 3rd, 2012.

Editor-in-Chief

Prof. (Dr.) Srikanta Patnaik
President, IRNet India
Intersceince Campus, Bhubaneswar
Mailto: iccsit.guwahati@gmail.com

&

Dr. Ramjeevan Singh Thakur,
Associate Professor,
Department of Computer Applications,
Maulana Azad National Institute of Technology, Bhopal-462051, MP, INDIA

Organized by:

About **ICCSIT 2012**

Computer Science and Information Technology have a profound influence on all branch of science, engineering, management as well. New technologies are constantly emerging, which are enabling applications in various domains and services. International Conference on Computer Science and Information Technology (CSIT) is organized by IRNet for the presentation of technological advancement and research results in the fields of theoretical, experimental, and applied area of Computer Science and Information Technology. CSIT aims to bring together developers, users, academicians and researchers in the information technology and computer science for sharing and exploring new areas of research and development and to discuss emerging issues faced by them. *Topics of interest for submission include, but are not limited to*:

- Algorithms
- Automated Software Engineering
- Bioinformatics and Scientific Computing
- Compilers and Interpreters
- Computer Animation
- Computer Architecture and Embedded Systems
- Computer Games
- Computer Graphics and Multimedia
- Computer Networks
- Computer Security
- Artificial Intelligence
- Bio-informatics
- Biomedical Engineering
- Computational Intelligence
- Computer Architecture & VLSI
- Computer Based Education
- Computer Graphics & Virtual Reality
- Computer Modeling
- Computer Networks and Data Communication
- Computer Simulation

ORGANIZING COMMITTEE

Chief patron/ Program Chair:

Prof. (Dr.) Srikanta Patnaik
President IRNet & Chairman, I.I.M.T.,
Bhubaneswar
Intersceince Campus,
At/Po.: Kantabada, Via-Janla, Dist-Khurda
Bhubaneswar, Pin:752024. Orissa, INDIA

Secretary IRNet:

Prof. Pradeep Kumar Mallick
IIMT, Bhubaneswar
Email:pradeepmallick84@gmail.com
Mobile No: 08895885152

Conference Coordinator:

Mr. Ajit Dash
IRNet , Bhubaneswar,India

Publication

Prof. Sushanta Kumar Panigrahi

Prof. Mritunjay Sharma

Prof. Sharada Prasad Sahoo.

Post Conference Coordinator

Bibhu Prasad Mohanty
Mob:08895995279

Head (System & Utilities)
Prof. Sanjay Sharma

Members:

Rashmi Ranjan Nath
Ujjayinee Swain
Bikash Kumar Rout
Pritika Mohanty

First Impression : 2012

Proceedings of International Conference on

COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

DISCLAIMER

The authors are solely responsible for the contents of the papers complied in this volume. The publishers or editors do not take any responsibility for the same in any manner. Errors, if any, are purely unintentional and readers are requested to communicate such errors to the editors or publishers to avoid discrepancies in future.

ISBN : 978-93-81693-22-3

Published by :

IPM PVT. LTD., Interscience Campus
At/PO.: Kantabada, Via: Janla, Dist: Khurda, Pin- 752054
Publisher's Website : www.interscience.in
E-mail: ipm.bbsr@gmail.com

Typeset & Printed by :

IPM PVT. LTD.

TABLE OF CONTENTS

Editorial

The mushrooming growth of the IT industry in the 21st century determines the pace of research and innovation across the globe. In a similar fashion Computer Science has acquired a path breaking trend by making a swift in a number of cross functional disciplines like Bio-Science, Health Science, Performance Engineering, Applied Behavioral Science, and Intelligence. It seems like the quest of Homo Sapience Community to integrate this world with a vision of Exchange of Knowledge and Culture is coming at the end. Apparently the quotation "Shrunken Earth, Shrinking Humanity" holds true as the connectivity and the flux of information remains on a simple command over an internet protocol address. Still there remains a substantial relativity in both the disciplines which underscores further extension of existing literature to augment the socio-economic relevancy of these two fields of study. The IT tycoon Microsoft addressing at the annual Worldwide Partner Conference in Los Angeles introduced Cloud ERP (Enterprise Resource Planning,) and updated CRM (Customer Relationship Management) software which emphasizes the ongoing research on capacity building of the Internal Business Process. It is worth mentioning here that Hewlett-Packard has been with flying colors with 4G touch pad removing comfort ability barriers with 2G and 3G. If we progress, the discussion will never limit because advancement is seamlessly flowing at the most efficient and state-of-the art universities and research labs like Laboratory for Advanced Systems Research, University of California. Unquestionably apex bodies like UNO, WTO and IBRD include these two disciplines in their millennium development agenda, realizing the aftermath of the various application projects like VSAT, POLNET, EDUSAT and many more. 'IT' has magnified the influence of knowledge management and congruently responding to social and industrial revolution.

The conference is designed to stimulate the young minds including Research Scholars, Academicians, and Practitioners to contribute their ideas, thoughts and nobility in these two integrated disciplines. Even a fraction of active participation deeply influences the magnanimity of this international event. I must acknowledge your response to this conference. I ought to convey that this conference is only a little step towards knowledge, network and relationship.

The conference is first of its kind and gets granted with lot of blessings. I wish all success to the paper presenters.

I congratulate the participants for getting selected at this conference. I extend heart full thanks to members of faculty from different institutions, research scholars, delegates, IRNet Family members, members of the technical and organizing committee. Above all I note the salutation towards the almighty.

Prof. (Dr.) Srikanta Patnaik
President, IRNet India and Chairman IIMT
Intersceince Campus, Bhubaneswar
Email: patnaik_srikanta@yahoo.co.in

&

Dr. Ramjeevan Singh Thakur,
Associate Professor,
Department of Computer Applications,
Maulana Azad National Institute of Technology, Bhopal-462051, MP, INDIA

Web Application to Obtain Interoperability Information

Ramyashree G & Tamilarasi T

ISE Department, MSRIT, Bangalore, India

Abstract - With the fast development of information technology, it is necessary that we should build a web application to obtain the product interoperability information. This paper describes a web application which enables to search for interoperability information about products. It acts as a consolidated repository, access and reporting application. This application maintains a central repository of products, and provides a user-friendly interface for the easy access of information and also provides web service so that other application or web services can also obtain the product interoperability information without the need of any user. After introducing the need for obtaining the product interoperability information, this paper explains the framework of the web application to obtain the product interoperability information.

Keywords - *Web application; Product interoperability information; Web services, Central repository, SOA.*

I. INTRODUCTION

In the modern days, with the availability of huge number and type of software and hardware products, obtaining the interoperability information i.e. "what works with what" has become a hot topic. Basic functionality that any computing system must provide is the retrieval of information from the database. With the growing complexity of information, and with the advent of distributed systems, it has become necessary to process any retrieval request coming from any system which is a part of the distributed environment. Fulfilling this requirement needs a generic retrieval system supporting interoperability which can formulate the request on the fly [1].

Interoperability is the ability to interconnect software products irrespective of their suppliers and vintages, to provide access to corporate data by any authorized user, and to maintain that interconnection and access over changes in suppliers and vintages [1]. The first step in achieving such a large interoperability is to follow similar development processes for collaborating domains, which provides homogeneity in their architectures. The second step would be to provide intra-domain and inter-domain semantic interoperability through proprietary and shared ontology systems [2].

A major challenge in interoperability among systems is interpretation of concepts from outside of one's domain of expertise. Therefore, the first task is to extract both data and services from participating application domains to allow systems to perform mutual business. The next task is to provide the means for communication of information and communication of meaning which are achieved through comprehensive and standard information and concept representations and communication through standard messages. Service oriented architecture is one of the major development techniques to achieve interoperability among the products [2].

Remaining section of the paper describes the frame work of web application which a client can use to obtain the interoperability information of the product; same technique can be extended to develop the web service to achieve the same functionality.

II. METHODOLOGY AND DESIGN

Software development methodology in software engineering is the framework that is used to structure, plan and control the process of developing an information system. Agile methodologies change the face of the software development by giving it a human touch (people over processes). Customers generally value ROI and time to market. Agile methodologies help in increasing ROI as well as reduce time to market (incremental working software).

Drilling down further, the Scrum flavor of agile software methodology is what is followed during the development phase. The scrum framework consists of practices and predefined roles: Product Owner represents the stakeholders and the business, Scrum Master maintains the processes, and the Team comprises a group of people working cross-functionally and is engaged in analysis, estimation, design, implementation, testing, and so on. Scrum consists of sprints, which span over two-three weeks. During the course of the sprint, daily effort is tracked using a tool called uTrack. This tool helps the scrum team to gather information in a burn down chart for each individual. The burn down chart is displayed to the team and provides a simple

view of the sprint progress. By using the information from uTrack, the individual productivity, effort variance, and scope variance can be calculated using simple formulae [4].

Good design is the key to effective engineering. Any design problem can be tackled in 3 stages: Study and understand the Problem, Identify gross features of at least one possible solution, Describe each abstraction used in the solution. Object oriented software development encourages viewing the problem as a system of cooperative objects. The objects incorporate both data and procedures. Object oriented approach has been used during the development. This approach has been chosen because of the advantages like, higher level of abstraction, seamless transition among different phases of software development, encouragement of good programming techniques, and promotion of reusability.

III. PARTIES IN WEBSERVICE

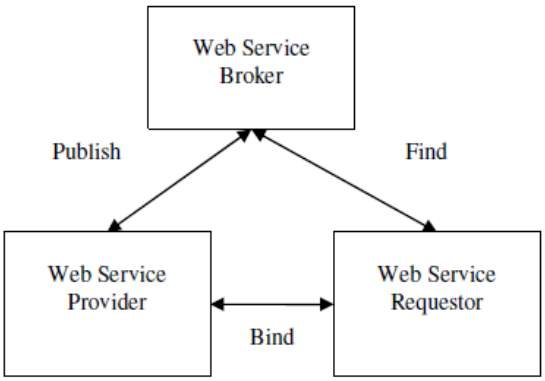


Fig. 1: Parties in Web Service [5]

Applications which can be used as the Web Services are called information services on Web. They are objects deployed on the Internet. Major model which has been used in the development of web services is Service Oriented Architecture (SOA) illustrated in Figure 1. In the framework, three parties involved in the life cycle of web service, which include service requestors (namely customers), service providers and service brokers.

Service providers publish the capabilities of their services, and rules to access them via Web Services Description Language (WSDL) descriptions, to these service brokers following the Universal Description, Discovery and Integration (UDDI) standards. Service brokers accept these services from service providers and classify them, at the same time they provide information services for service requestors. Customers request services from service platforms. Services are located based on possibly registry-specific UDDI taxonomy. The three roles interact at the web service platforms, which make service composition and execution more convenient.

The core of Web Service is XML. Its protocols are composed of SOAP(Simple Object Access Protocol), WSDL(Web Service Description Language) and UDDI(Universal Description Discovery and Integration). XML provides uniform data format for Web services. Description of information, service and workflow all adopt XML as their definition language. SOAP is a light protocol, which is used for exchanging and encoding the information of XML.

By using HTTP, SMTP and FTP protocols, it can be compatible with existing communications. By using XML, WSDL describes one port or an information service, which is used to define Web services and its invoking method. UDDI provides a framework, which describes and finds commercial service on Web. UDDI is also the standard and criterion of service brokers [5].

IV. WEB TOOL ARCHITECTURE

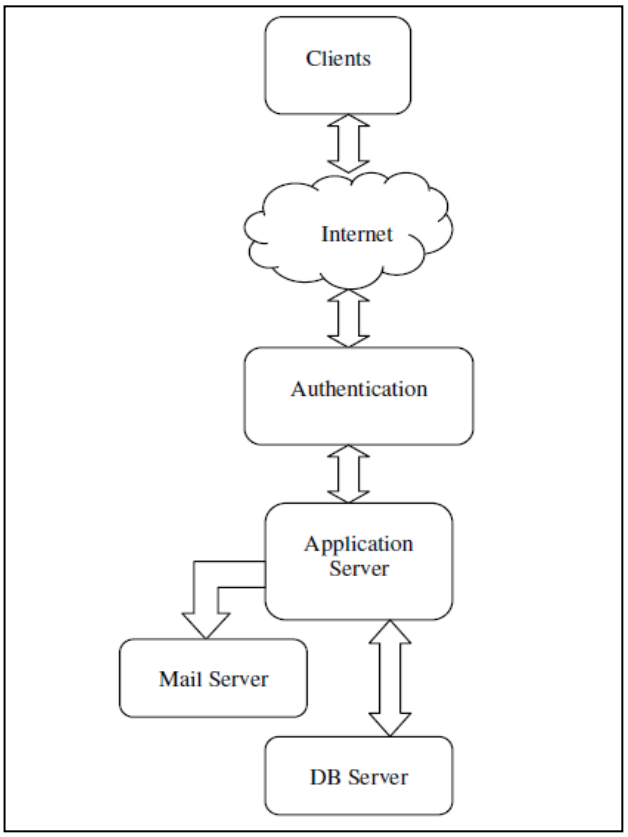


Fig. 2 : Network Diagram

Clients who want to get the particular product interoperability information with the other products can get the required information by sending the request to the application. In single request itself user can obtain

the interoperability information of multiple products. Requests from clients to obtain product interoperability information will be forwarded to the application server, after authenticating the client. Application server will then interacts with database server and retrieves the interoperability information of the products that are been requested by the client. Database server will be connected to a central repository which contains all the information about software and hardware products. Application server will also interacts with mail server in order to send automatic mails to the requestor. This interaction has been depicted in the figure [2].

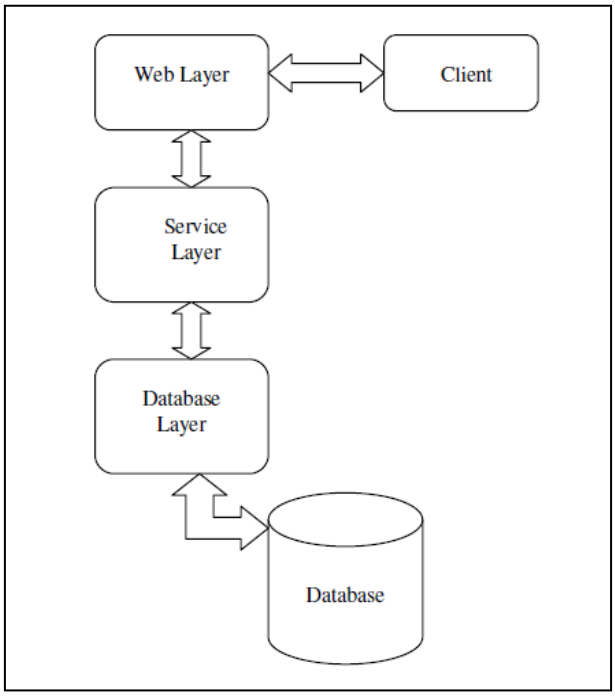


Fig. 3 : Application Framework

Client starts the interaction with the application in the web layer. In web layer, there will be a service to authenticate the client. After authentication, web layer provides a user friendly interface which allows the client to interact with the application. Client can raise the request for product interoperability information using the interface. Request from the client will be forwarded to the service layer as an http request; this request will then be forwarded to the database through database layer as pojo's. Database layer will then query the database using stored procedures. The result pf this query will be sent back to web layer through service layer as pojo's. From web layer the interoperability information will be forwarded to the client in the form of http response. The entire process is depicted in figure [3].

V. WEB SERVICE FRAMEWORK

In service resource layer, there will be different web services which give information about software products and hardware products. These services will give the information like solution name, version and compatibility information about the product. These services will also give property information which has been associated with the hardware products. Web services in service resource layer obtain the required information by accessing the central repository, which has information about the software and hardware products.

Application layer will contain the clients and other web services, which want to obtain the interoperability information. Service layer has service catalog, which uses different web services like authentication, registration and finding. All these web services collaborate with each other in order to forward the request from the requestor to obtain the interoperability information to the service resource layer.

Each service of the service resources layer is registered in the center with the PUBLISH port. The information of service types and service instance location is enrolled into the service registration catalog. The application of the application layer finds appropriate service instance location with the FIND port according to the information of service types, and then the application layer can be bound with the service instance and invokes different functions of the service [3]. Figure [4] shows the web service framework. The services are generated using top down approach where a java program has been written using springs, apache cxf tool, while executing the program a WSDL from the URL of the executed program will be obtained. Once the WSDL is obtained, one can create a client or run the requests from SOAP UI tool by using this WSDL file.

VI. CONCLUSION

The tool allows the clients to query it in order to obtain different hardware and software products interoperability information. Using the tool clients can obtain the information which tells which version of the product can inter operate with which version of the other product, irrespective of whether it is a hardware or software product. The concept used for the development of the tool can be extended to build web service, using which other application can obtain the interoperability information, without need for the user to query the tool.

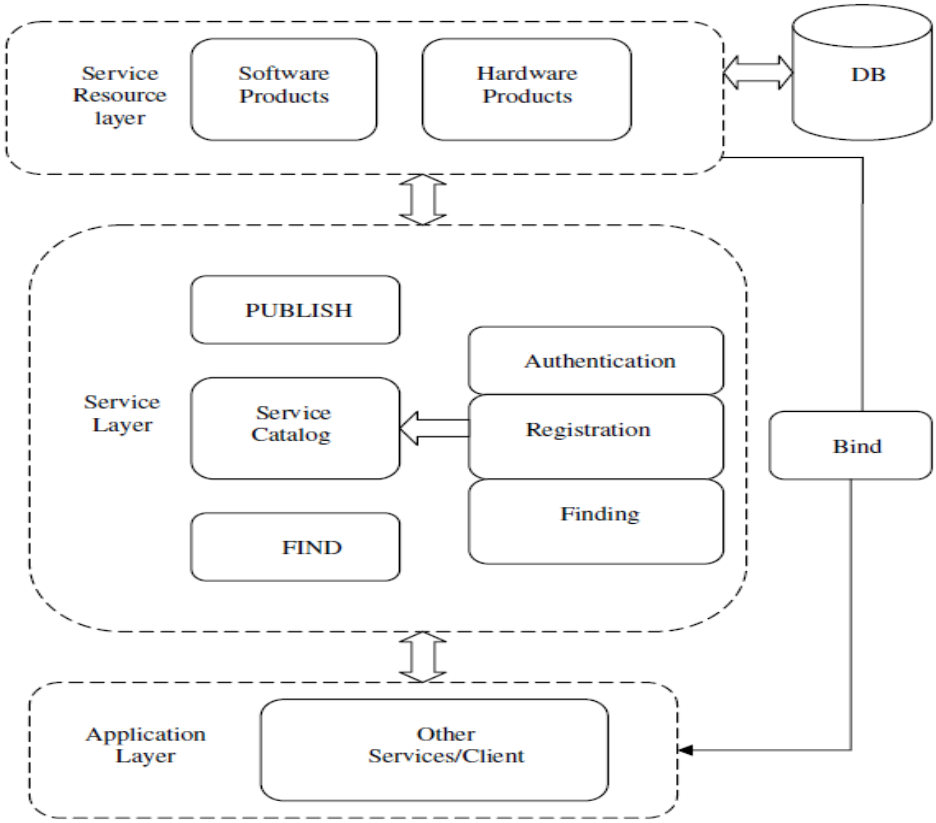


Fig. 4 : Web Service Framework [3]

VII. ACKNOWLEDGMENT

My hearty thanks to my guide Tamilarasi, for the encouragement, support and guidance from the beginning of the project till the completion of it.

I offer my regards and blessings for all those who have support me during the project period, for better understanding of the subject, which helped in the completion of the project on time.

REFERENCES

[1] Debajyoti Mukhopadhyay., "A Generic Information Retrieval System to Support Interoperability."

[2] Kamran Sartipi., Azin Dehmoobad Cross., "Domain Information and Service Interoperability"., Proceedings of iiWAS., 2008.

[3] Haigang An., Ning Liu., "Research on Information Services of Sustainable Utilization of Natural Resources Based on Web"., 2008 IEEE.

[4] Agile Methodology http://en.wikipedia.org/ wiki/Agile_Methods.

[5] WebServiceshttp://www.w3schools.com.

◈◈◈

Atherosclerosis Detection Using Music

Niharika Jha

Department of Inter-Disciplinary Science and Technology, I2IT, Pune, India

Abstract - The recent advancements in the field of embedded electronics have given birth to a number of bio-medical consumer devices which include portable blood pressure monitors and blood glucose monitors. Atherosclerosis and cardiovascular disease take a huge toll on our society. Atherosclerosis typically begins in early adolescence, and is found in major arteries, yet is asymptomatic and not detected by most diagnostic methods during life. However it does produce a noticeable change in the waveform generated by an ECG. It is known that ECG interpretation requires a qualified individual, but if the irregularities observed in the ECG signal are used to bring about audible or visual changes, it would enable a common man to diagnose his condition before it gets critical. This project aims at using irregularities in an ECG signal to induce tempo in a polyphonic music piece using the duration of the QRS complex. The amalgamation of a music player and an ECG is yet to be seen in the market. Implementation of this project will also bring forth the feasibility of such a product to see how it will benefit the patient.

Keywords - *Atherosclerosis, Electrocardiogram, ECG reading interpretation ,Power-line interference, audio tempo changing.*

I. INTRODUCTION

Coronary artery disease (CAD) affects more than 16 million people, making it the most common form of heart disease. CAD and its complications, like arrhythmia, angina pectoris, and heart attack (also called myocardial infarction), are the leading causes of death. CAD most often results from a condition known as atherosclerosis, which happens when a waxy substance forms inside the arteries that supply blood to your heart.

Major symptoms :

Atherosclerosis may be present for years without causing symptoms. This slow disease process can begin in childhood. In some people, the condition can cause symptoms by the time they reach their 30s. In others, they do not have symptoms until they reach their 50s or 60s. But, as the blockage gets worse, the slowed blood supply to the heart may begin to cause something called angina pectoris. Patients often say that angina is like a squeezing, suffocating, or burning feeling in their chest. The pain usually happens when the heart has an extra demand for blood, like during exercise or times of emotional stress.

Angina tends to start in the center of the chest but may move to your arm, neck, back, throat, or jaw. Some people say they feel numbness or a loss of sensation in their arms, shoulders, or wrists. An episode usually lasts no more than a few minutes and goes away with rest.

Diagnosis:

A baseline electrocardiogram (ECG or EKG), which records your heart's electrical activity while you sit quietly. An exercise ECG, also known as a stress test, will show how your heart responds to increasing exercise. Both tests are designed to show if your heart is not working properly, most likely due to a lack of oxygen.

An exercise thallium test, also called a nuclear stress test, which uses a radioactive substance that is injected into your bloodstream to show how blood flows through your arteries. Doctors can see if your heart muscle is damaged or dead, or if you have a serious narrowing in an artery. For people who cannot take an exercise test, medicines can be given that make your heart beat as if you were exercising.

Echocardiography, which uses sound waves to produce an image of the heart to see how it is working.

Coronary angiography, which is performed in the cardiac catheterization laboratory. After you are given medicine to relax you, dye is injected into your bloodstream to give doctors an x-ray "movie" of heart action and blood flow through your valves and arteries (called an angiogram). Doctors can see the number of blockages that you have and how serious those blockages are. Doctors often use this test to find out which treatment option may be best for you.

Positron emission tomography (PET) scanning, which uses information about the energy of certain elements in your body to show whether parts of the

heart muscle are alive and working. A PET scan can also show if your heart is getting enough blood in order to keep the muscle healthy[1].

Changes brought about in an ECG rhythm:

Normal sinus rhythm

- Rhythm - Regular
- Rate - (60-100 bpm)
- QRS Duration - Normal
- P Wave - Visible before each QRS complex
- P-R Interval - Normal (<5 small Squares. Anything above and this would be 1st degree block)
- Indicates that the electrical signal is generated by the sinus node and travelling in a normal fashion in the heart.

Heart blockages occur when electrical signals that pump blood in and out of the ventricles are blocked.Heart blockages are detected by an EKG exam. There are three types of heart blockage: first, second, and third degree. First degree is less severe and third degree is the most severe. Mobitz type I and II are categorized separately, but they are forms of second degree blockage.

1. First Degree

First degree heart blockage occurs when electrical impulses are slowed as they travel down the atrium to the ventricles. First degree blockages does not exhibit symptoms, and it's more common among young, active people. Young people have more active vagus nerves, and this large heart nerve inhibits electrical activity in heart cells.

- Rhythm - Regular
- Rate - Normal
- QRS Duration - Normal
- P Wave - Ratio 1:1
- P Wave rate - Normal
- P-R Interval - Prolonged (>5 small squares)

2. Second Degree

Second degree blockage is more serious than first degree. This condition is caused when electrical activity is slowed so badly that they do not reach the ends of the ventricles. This inhibits proper pumping of blood. Second degree blockage is further divided into categories.

Mobitz Type I

Mobitz type I is a type of second degree blockage where the electrical activity becomes weaker and weaker until the heart skips a beat. The process is continued consistently, so blood does not get pumped properly. The decreased heart rate causes tissue to lose oxygen from lower blood circulation. The main symptom of mobitz type I is dizziness. The age group at high risk of mobitz type I is the elderly, but congenital heart defects can be passed from the mother to the infant.

- Rhythm - Regularly irregular
- Rate - Normal or Slow
- QRS Duration - Normal
- P Wave - Ratio 1:1 for 2,3 or 4 cycles then 1:0.
- P Wave rate - Normal but faster than QRS rate
- P-R Interval - Progressive lengthening of P-R interval until a QRS complex is dropped

Mobitz Type II

Mobitz type II is a more serious condition where the electrical activity in the heart is irregular. In some contractions, the heart beats regularly. In other contractions, the electrical signals are blocked and the heart skips a beat. This type of condition is remedied using a pacemaker. The age group at high risk of mobitz type II is the elderly, but younger patients with heart disease also suffer from mobitz type II conditions.

- Rhythm - Regular
- Rate - Normal or Slow
- QRS Duration - Prolonged
- P Wave - Ratio 2:1, 3:1
- P Wave rate - Normal but faster than QRS rate
- P-R Interval - Normal or prolonged but constant

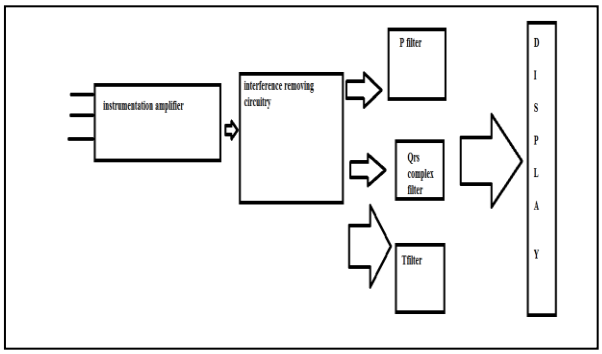
3. Third Degree

Third degree heart blockage is when whole parts of the ventricles do not receive electrical impulses. This causes irregular heart beats and improper blood regulation. Third degree heart blockage can lead to cardiac arrest if not treated immediately. Some doctors insert temporary pacemakers until a permanent one can be placed.[2].

- Rhythm - Regular
- Rate - Slow
- QRS Duration - Prolonged
- P Wave - Unrelated
- P Wave rate - Normal but faster than QRS rate
- P-R Interval - Variation
- Complete AV block. No atrial impulses pass through the atrioventricular node and the ventricles generate their own rhythm.[3].

As seen above the nature of the ECG waveforms for different kinds of heart blockages are different and hence the parameters such as QRS duration, Pwaveratio,P-R interval can be used for tempo changing to detect bloackages.

II. BLOCK DIAGRAM OF AN ECG.



This is a simplified block diagram of an ECG.

It shows that 3 leads are at least required to properly interpret an ECG signal. One lead attached to the right leg acts as the reference or the ground signal. The 3 signals are amplified by an instrumentation amplifier and sent to the filters. The band pass filters are used for separation of the three characteristic waves and these waves are now sent to the display section.

The recommended system bandwidth is between 0.05 and 150 Hz. Of great importance in ECG diagnosis is the low-frequency response of the system, because shifts in some of the low-frequency regions, e.g., the ST segment, have critical diagnosis value. While the heart rate may only have a 1-Hz fundamental frequency, the phase responses of typical analog high-pass filters are such that the system corner frequency must be much smaller than the 3-dB corner frequency where only the amplitude response is considered. The system gain depends on the total system design. The typical ECG amplitude is ±2 mV, and if A/D conversion is used in a digital system, the enough gain to span the full range of the A/D converter is appropriate.

To first obtain an ECG the patient must be physically connected to the amplifier front end. The patient amplifier interface is formed by a special bio electrode that converts the ionic current flow of the body to the electron flow of the metallic wire. These electrodes typically rely on a chemical paste or gel with a high ionic concentration. This acts as the transducer at the tissue-electrode interface. For short-term applications, silver-coated suction electrodes or “sticky” metallic foil electrodes are used. Long-termrecordings, such as for the monitored patient, require a stable electrode-tissue interface, and specialadhesive tape material surrounds the gel and an Ag+/A g+Cl– electrode.

At any given time, the patient may be connected to a variety of devices, e.g., respirator, blood pressure monitor, temporary pacemaker, etc., some of which will invade the body and provide a low-resistance pathway to the heart. It is essential that the device not act as a current source and inject the patient with enough current to stimulate the heart and cause it to fibrillate. Some bias currents are unavoidable for the system input stage, and recommendations are that these leakage currents be less than 10 μA per device. This applies to the normal setting, but if a fault condition arises whereby the patient comes in contact with the high-voltage side of the alternating current (ac) power lines, then the isolation must be adequate to prevent 10 μA of fault current as well. This mandates that the ECG reference ground not be connected physically to the low side of the ac power line or its third-wire ground. For ECG machines, the solution has typically been to AM modulate a medium frequency carrier signal (≈400 kHz) and use an isolation transformer with subsequent demodulation. Other methods of signal isolation can be used, but the primary reason for the isolation is to keep the patient from being part of the ac circuit in the case of a patient-to-power-line fault. In addition, with many devices connected in a patient monitoring situation, it is possible that ground loop currents will be generated. To obviate this potential hazard, alow-impedance ground buss is often installed in these rooms, and each device chassis will have an external ground wire connected to the buss[3].

III. ECG SIGNAL ACQUISITION:

In clinical practice the standard 12 lead ECG obtained using four limb leads and chest leads in 6 positions. The left and right arm and the left leg(I,II and III respectively) are used as reference for chest leads. The augmented limb leads known as aVR, aVL and aVF are obtained by using the exploring electrode on the limb indicated by the lead name, with reference being Wilsons central terminal. The hypothetical equilateral triangle formed by leads I,II, and III is known as

Einthoven's triangle. The center of the triangle represents Wilsons central terminal. The six chest leads V1-V6 are obtained from six standardized positions on the chest.

Some standard important features of the clinical ECG are :

1) Peak to peak value about 2mV.
2) Bandwidth-.5 – 50 Hz
3) Sampling rate- 500Hz[4].

IV. REMOVAL OF ARTIFACTS:

The different kinds of interference waveforms (artifacts) added to the ECG signal during the recording are:

1) EMG related to coughing, breathing, or squirming affecting the ECG.
2) Breath, lung, or bowel sounds contaminating the heart sounds(PCG).
3) Muscle sound (VMG) interference in joint sounds (VAG).
4) Maternal ECG getting added to the fetal ECG of interest.
5) Electrical interference external to the subject and recording system.
6) High-frequency noise in the ECG.
7) Motion artifact in the ECG.
8) Noise due to variation of electrode skin contact impedance.
9) Power-line Interference in ECG signals.
10) Noise generated by electronic devices used in signal processing circuits.

This paper concentrates mainly on removal of power line interference . However the final product will also ensure that the noise generated by the electronic devices used in the entire circuit do not corrupt the ECG signal.

There are mainly three approaches to remove noise and interference,

(1) Frequency-domain filtering (Notch Filter)

(2) Optimal (Wiener) filtering,

(3) Adaptive filtering

Notch Filter :

It is well known or simplest filter to remove the power line interface notch filter compute the Fourier transform of the signal delete undesired component and the inverse Fourier transform. There are two methods for implementation of the notch filter .First one is remove the artifact or set its value to zero. In second method the 50Hz artifact set to be average value of the signal. Later methods not remove the 50Hz component of the signal , but noise removing performance is average and in first one filter removes the 50Hz component of the ECG signal. After closely examining an IIR notch filter and three Type 1 FIR band-stop filters of varying order, it was found the IIR to perform best overall. Although the IIR filter's phase response is non-linear, almost all of the non-linearities occur within the stop-band. This would seem to indicate that it's shifting the phase of frequencies we're not interested in anyway. The IIR's low computation cost is also of importance especially when we are looking at implementing some sort of noise filter for an actual piece of medical equipment. This implies finite computational resources and keeping costs down. The IIR filter achieves both of these goalswhile still delivering a high quality filtered signal.[5]

Wiener Filter :

The notch filter and other pass band, band stop filters are fixed filter, they use only limited resources or we cannot change its performance according to our need. Wiener filter use the statistical characteristics for noise removing process like reference signal or secondary recorded ECG signal.

We can change its parameter to get the optimal results, so then we also called it optimal filter . Wiener filter theory provides for optimal filtering by taking into account the statistical characteristics of the signal and noise processes. The filter parameters are optimized with reference to a performance criterion, the output is guaranteed to be the best achievable result under the conditions imposed and the information provided.

Adaptive Filtering :

Adaptive filters are self-designing filters based on an algorithm which allows the filter to "learn" the initial input statistics and to track them, if they are time varying. These filters estimate the deterministic signal and remove the noise uncorrelated with the deterministic signal , we have considered adaptive impulse correlated filter which requires the signal and a reference input. The least mean square algorithm is used to adjust the weights of the adaptive filter in order to minimize the error and estimate the deterministic component through filter output.

The LMS adaptive filter is widely used to filter the ECG signal, but the existing LMS adaptive filters adapt to the environment showing limitations in the given filter, so its convergence and performance cause distortions and even poor performance, depending on the environment and the patient's condition. A Dynamic

Structure Adaptive Filter as proposed byJu-Won Lee and Gun-Ki Lee[6].

V. PEAK DETECTION:

There are many methods that have been implemented for detection of P,QRS,T peaks over the years, two of the most famous ones being 1)derivative based and 2) Pan Tompkins algorithm.

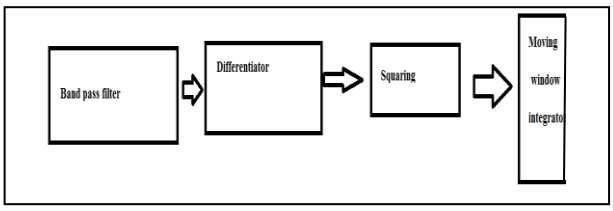
The algorithm developed by Pan and Tompkins identifies QRS complexes based on analysisof the slope, amplitude, and width of the QRS. The various stages of the algorithm are :-

Band pass filter: - The bandpass filter, formed using lowpass and highpass filters, reduces noise in the ECG signal. Noise such as muscle noise, 60 Hz interference, and baseline drift are removed by bandpass filtering. The signal is then passed through a differentiator to provide a large response at the high slopes that distinguish QRS complexes from low-frequency ECG components such as the P and T waves.

Differentiator:- The signal is then passed through a differentiator to provide a large response at the high slopes that distinguish QRS complexes from low-frequency ECG components such as the P and T waves.

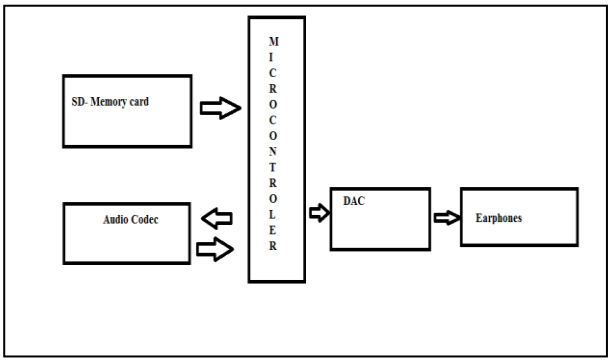
Squaring :-The next operation is the squaring operation, which emphasizes the higher values expected due to QRS complexes and suppresses smaller values related to the P and T waves

Moving-window Integrator:- The squared signal is then passed through a moving-window integrator of window length N = 30 samples (for the sampling frequency of fs = 200 Hz). The expected result is a single smooth peak related to the QRS complex for each ECG cycle. The output of the moving-window integrator may be used to detect QRS complexes, measure RR intervals, and determine the duration of the QRS complex.[4].



VI. BLOCK DIAGRAM OF A MUSIC PLAYER:

A music player consists of a memory unit in this case a micro SD card interfaced to a controller. The controller reads the memory device and sends data to the audio codec unit. The audio codec unit converts an MP3 file to a wav file which can be sent to the earphones through a DAC.



VII. INDUCTION OF TEMPO IN MUSIC

The tempo refers to the pace of a musical excerpt. Given a metrical structure, tempo is defined as the rate of the beats at a given metrical level, for example the quarter note level in the score. It is inversely proportional to the pulse period. For music with almost constant tempo, tempo induction is feasible with around 80% accuracy and a relatively good robustness to distortion. AnssiKlapuri from the Tampere University of Technology submitted one algorithm as a GNU/Linux binary, referred to as Klapuri algorithm.The onset time is defined as the beginning time of abeat or note played. To track a beat, spectrogram analysis to raw signals with 4096window size is applied.

The spectrogram represents the power of different frequencies at different time indices. Let *p(t,f)* being the spectrogram of given signal. The degree of onset *d(t,f)* is given by

$$d(t,f) = p(t,f) - pp + \max(0, p(t+1,f) - p(t,f))$$

Where $pp = \max(p(t-1,f), p(t-1,f \pm 1), p(t-2,f))$,

Finally, the degree of onset is a function of time and given by

$$D(t) = \Sigma f\, d(t,f)$$

In this algorithm, possible inter-onset interval is looped over from 9 to 120 in unit of the parameter of *D(t).*

Finally, normalize the score and obtain the IOI reliability score.

An important aspect of this algorithm lies in the feature list creation block: the differentials ofthe loudness in 36 frequency sub-bands are combined into 4 “accent bands”, measuring the “degree of musical accentuation as a function of time.” The goal in this procedure is to account for subtle energy changes that might occur in narrow frequency sub-bands as well as wide-band energy changes. The pulse induction block implements a bank of comb filters.

Another particularity of this algorithm is the joint determination of three metrical levels (thetatum, the

tactus and the measure) through probabilistic modeling of their relationships and temporal evolutions. After computing the tactus beats of the whole test excerpt, the tempo was computedas the median of the IOIs of the excerpt's latter half.

VIII. CONCLUSION :

The above paper is an intermediate step to realizing the feasibility of changing the tempo of music in accordance with an ECG signal. When achieved, the product will have fixed values of the tempo at which the music should be played for every kind of heart blockage.

REFERENCE

[1] http://texasheart.org/HIC/Topics/Cond/ CoronaryArteryDisease.cfm

[2] http://www.ambulancetechnicianstudy.co.uk/ rhythms.html

[3] http://www.cisl.columbia.edu/kinget_group/ student_ projects/ECG% 20Report/E6001% 20ECG%20final%20report.htm

[4] Rangaraj M. Rangayyan" Bio-medical signal analysis:A case study approach." WILEY interscience.

[5] Ju-Won Lee and Gun-Ki Lee Design of an Adaptive Filter with a Dynamic Structurefor ECG Signal Processing, International Journal of Control, Automation, and Systems, vol. 3, no. 1, pp. 137-142, March

[6] YatindraKumar,Assistant ProfessorDepartment of EEE,G.B.PantEnggCollege,Pauri,(U.K)India; Gorav Kumar Malik, Research Scholar, Department of EEE ,G.B.PantEnggCollege,Pauri (U.K)IndiaPerformance Analysis of different Filters for Power LineInterface Reductionin ECG Signal International Journal of Computer Applications (0975 – 8887) Volume 3 – No.7, June 2010

[8] Fabien Gouyon Audio Features for the Computation Ofrhythm Periodicity Functions and their use in Tempo Induction and Music Content Processing 2005

◈◈◈

Combining Server And Storage Virtualization: A New Dimension for SMB's

Mohammad Arif Baig

Advance Networking and Telecom Dept (SOIT), International Institute of Information Technology, Pune, India

Abstract - The virtualization wave is quickly reaching its way down into the small-to-medium-sized business. Virtualization provides unmatched flexibility, performance, and utilization by allowing you to move server workloads from one virtual workspace to the next, maximizing server resources on the fly based on your business needs. Server virtualization eliminates the conventional, one application per server model and allows businesses to run multiple, virtual servers on a single physical machine. Storage virtualization helps the storage administrator perform the tasks of backup, archiving, and recovery more easily, and in less time, by disguising the actual complexity of the SAN. Storage and server virtualization are complementary technologies that helps to build a completely virtualized infra-structure. When used together, server and storage virtualization are intended to derive greater benefit from each technology than deployed alone.

Keywords - *SAN, NAS, SMB, ILM, HBA, iSCSI.*

I. INTRODUCTION

A. What is Virtualization?

Virtualization is a method of running multiple independent virtual operating systems on a single physical computer. It is a way of maximizing physical resources to maximize the investment in hardware. Since Moore's law has accurately predicted the exponential growth of computing power and hardware requirements for the most part have not changed to accomplish the same computing tasks, it is now feasible to turn a very inexpensive 1U dual-socket dual-core commodity server into eight or even 16 virtual servers that run 16 virtual operating systems. Virtualization technology is a way of achieving higher server density. However, it does not actually increase total computing power; it decreases it slightly because of overhead.

Virtualization is being used by a growing number of organizations to reduce power consumption and air conditioning needs and trim the building space and land requirements that have always been associated with server farm growth. Virtualization also provides high availability for critical applications, and streamlines application deployment and migrations. Virtualization can simplify IT operations and allow IT organizations to respond faster to changing business demands.

B. When to use virtualization

Virtualization is the perfect solution for applications that are meant for small- to medium-scale usage. Virtualization should not be used for high-performance applications where one or more servers need to be clustered together to meet performance requirements of a single application because the added overhead and complexity would only reduce performance, e.g. We're essentially taking a 12 GHz server (four cores times three GHz) and chopping it up into 16 750 MHz servers. But if eight of those servers are in off-peak or idle mode, the remaining eight servers will have nearly 1.5 GHz available to them [1].

C. Advantages of Using Virtualization

Today's IT intensive enterprise must always be on the lookout for the latest technologies that allow businesses to run with fewer resources while providing the infrastructure to meet today and future customer needs. Virtualization utilizing Intel Virtualization Technology is the cutting edge of enterprise information technology. Intel is closely working with VMware, XENSource, Jaluna, Parallels, tenAsys, VirtualIron, RedHat, Novell and other VMM developers.

C.1. Server Consolidation

It is not unusual to achieve 10:1 virtual to physical machine consolidation. This means that ten server applications can be run on a single machine that had required as many physical computers to provide the unique operating system and technical specification environments in order to operate. Server utilization is optimized and legacy software can maintain old OS configurations while new applications are running in VMs with updated platforms.

C.2. Testing and development

Use of a VM enables rapid deployment by isolating the application in a known and controlled environment.

Unknown factors such as mixed libraries caused by numerous installs can be eliminated. Severe crashes that required hours of reinstallation now take moments by simply copying a virtual image.

C.3. Dynamic Load Balancing and Disaster Recovery

As server workloads vary, virtualization provides the ability for virtual machines that are over utilizing the resources of a server to be moved to underutilized servers. This dynamic load balancing creates efficient utilization of server resources. Disaster recovery is a critical component for IT, as system crashes can create huge economic losses. Virtualization technology enables a virtual image on a machine to be instantly re-imaged on another server if a machine failure occurs.

C.4. Virtual Desktops

Multinational flexibility provides seamless transitions between different operating systems on a single machine reducing desktop footprint and hardware expenditure.

C.5. Improved System Reliability and Security

Virtualization of systems helps prevent system crashes due to memory corruption caused by software like device drivers. VT-d for Directed I/O Architecture provides methods to better control system devices by defining the architecture for DMA and interrupt remapping to ensure improved isolation of I/O resources for greater reliability, security, and availability [2].

II. SERVER VIRTUALIZATION

Server virtualization is the partitioning of a physical server into smaller virtual servers. In server virtualization the resources of the server itself are hidden, or masked, from users, and software is used to divide the physical server into multiple virtual environments, called virtual or private servers.

One common usage of this technology is in Web servers. Virtual Web servers are a very popular way of providing low-cost web hosting services. Instead of requiring a separate computer for each server, dozens of virtual servers can co-reside on the same computer.

Virtualization can drastically reduce the number of servers in a data center, thus decreasing electricity consumption and waste heat, and consequently the size of the necessary cooling equipment. Some investment in software and hardware may be required to implement virtualization, but it is usually modest compared to the savings achieved. There are three popular approaches to server virtualization: the virtual machine model, the paravirtual machine model, and virtualization at the operating system (OS) layer.

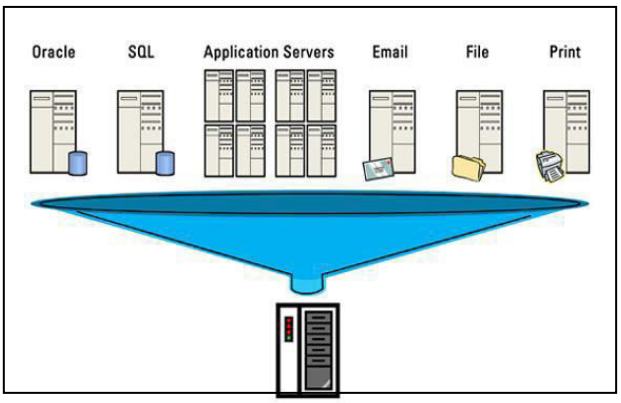


Fig. 1 : Virtualization of Servers

Benefits of Server Virtualization

Advances in Server Virtualization technology continue to highlight the many benefits of using server virtualization to efficiently consolidate servers, save physical space, provide more flexibility, and manage disaster recovery.

A. Consolidate Servers

The most benefit of server virtualization is the capability to consolidate applications on existing servers. Business applications usually don't require anywhere close to the resource capacity of their computers—perhaps an average of 20%. Several applications can be run on shared servers using virtual environments, allowing up to a 60-80% [3] resource utilization of a company's servers. The need for fewer servers results in significant savings in new server hardware spending, and additional savings are realized in lower maintenance costs for servers.

B. Conserve Physical Space

Using fewer servers also reduces the need for additional physical data center space. Limiting the size of the physical footprint means cost savings in heating, cooling, and electricity, as well as other maintenance needs for a facility.

C. Provide Flexibility

In the constantly changing IT environment, flexibility is crucial to maintaining older applications as well as developing and testing new applications. Different operating systems may also be required on the same hardware platform. Server Virtualization can allow a standard virtual server that can be easily duplicated to speed up server deployment, and can provide the environment to run legacy systems along with the newer versions of applications. Programs in development can be easily tested in virtual environments, and the migration of applications can be accomplished without interrupting business.

D. Manage Disaster Recovery

Planning for better disaster recovery management is another benefit of server virtualization technology. Traditional tape backup systems have typically one or two days for a complete restoration. Virtualization could allow a restoration in less than a half a day because the systems and applications already exist in other untouched virtual environments and just need to be brought online if proper planning, updating, and testing have been done. If a system fails, it can be automatically switched to a standby server or brought back to its normal state with a virtual copy of the original image. Virtualization strategies vary, but the benefits continue to be impressive in reducing hardware requirements and physical space allotments, as well as allowing the maximum strategic flexibility for allocating resource and planning for efficient disaster recovery [4].

III. STORAGE VIRLUALIZATION

In its simplest form, storage virtualization pools all of your storage resources into a single logical entity; this should sound familiar since it's the same process you went through when you started virtualizing your servers.

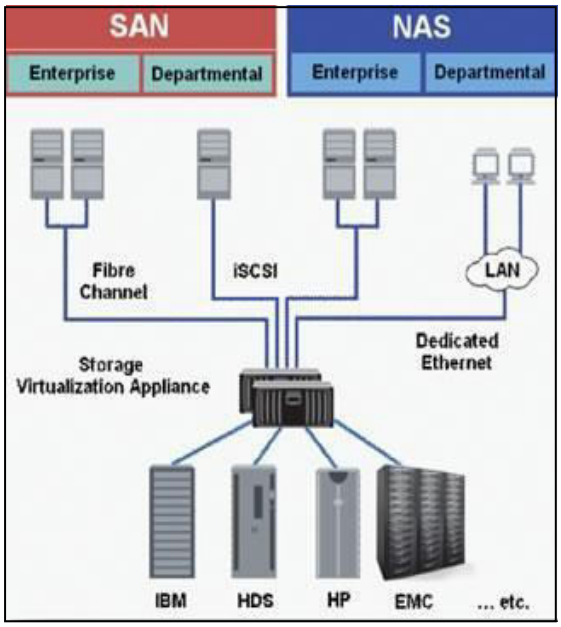


Fig. 2 : Storage Virtualization

Storage virtualization is an abstraction that presents servers and applications a view of storage that is different from that of actual physical storage, typically by aggregating multiple storage devices and allowing them to be managed in one administrative console.

Benefits of Storage Virtualization

A. Single point of administration:

Customers understand that a little friendly competition between storage vendors can help reduce the price of storage; SAN virtualization can be implemented in multi-vendor storage environments, so if a potential customer doesn't run the brand of storage equipment that you offer, converting them to a SAN virtualization setup could open the door to storage hardware sales. SAN-based storage virtualization provides this benefit by virtue of the fact that all of the SAN storage is provisioned to the virtualization device, so from that point on, all of the storage administration occurs at the virtualization layer.

B. Non-disruptive data migration:

Many customers do not replace their storage arrays when the lease or support expires or when the products are fully depreciated -- not because it is cheaper to keep the storage (the manufacturers make sure it's not), but because it is very difficult to migrate to the next storage platform. Without storage virtualization, migrations often require application outages and lots of sweat equity. These efforts are extremely difficult to coordinate across the various teams and business units involved. With SAN virtualization, the storage team can execute disk array swap-outs without impacting anyone else. This capability brings benefits to storage managers and VARs alike; migrations suddenly become not only possible but easy enough to make them worthwhile, allowing storage managers to take advantage of the declining cost of storage and bringing VARs an opportunity to sell new storage.

C. Information lifecycle management (ILM):

Customers want to know that they are putting their application data on the most appropriate tier of storage. Data access patterns are a key criterion in determining where to put the data. Often, a single business application or database has certain regions of data that are frequently accessed and require high-performance storage, while other regions are rarely accessed and could exist on more cost-effective storage. Because SAN virtualization tools sit between the server and the storage hardware, they have awareness of the access patterns. If the virtualization engine can use this access pattern information and leverage its online migration features, it may be possible to transparently relocate frequently accessed data to more expensive, high-performance storage and move less frequently accessed data to less expensive storage, bringing true ILM within reach.

D. Improved allocation efficiencies

Storage managers know that improving asset utilization is a quick way to lower the total cost of ownership (TCO) for their department. One of the common causes of low utilization is that the application teams demand more storage than they need. This may be because the process for requesting more storage is too slow; on the other hand, it could be because the application is new and there's not enough history to properly plan for growth. Storage virtualization promises to solve both problems. In the first case, the pace of deployment can be improved when all storage, regardless of brand or type, has a single administrative interface for allocation. And capacity planning challenges could be alleviated with thin provisioning services in the virtualization layer, which allow pre-allocation of storage and shared free space across applications optimizing unused disk, which is the most expensive storage asset.

E. *Heterogeneous replication*

One of the huge challenges associated with maintaining agnosticism among disk array vendors is disaster recovery replication. Most array-based storage replication is not heterogeneous, meaning that the production and disaster recovery frames must be of the same brand and often of the same type. Host-based replication options are heterogeneous, but management is cumbersome when a large number of hosts have replicated data. SAN virtualization can split the difference, providing a single method of replication for multiple types of storage arrays and a limited number of management points [5].

IV. INTEGRATING STORAGE AND SERVER VIRTUALIZATION FOR SMB's

Server virtualization projects are no longer being implemented in the largest data centers. The virtualization wave is quickly reaching its way down into the small-to-medium-sized business. In fact, the payoff for the SMB may be even greater than in the larger enterprise. SMBs, however, have typically had one significant disadvantage to the larger enterprise, accessibility to shared storage. Shared storage is the key in how the SMB can achieve server virtualization's full potential.

One of the prime focuses in any virtualization project, whether it's storage or server and especially both, is to conduct an inventory of the servers, storage devices and such that will be involved. This includes things such as the host bus adapters (HBA) and storage area network (SAN) switches, and the software and firmware revisions.

Check the hardware compatibility lists (HCL) for both virtualization products and make sure your configuration conforms. This is getting easier as virtualization vendors work to make their products interoperable. For example, VMware Inc., now owned by EMC Corp., is aggressively promoting its VMware Infrastructure 3, which ties VMware's ESX Server 3 and related products with storage virtualization, and associated hardware and software. Recently, both Emulex Corp. and QLogic Corp. announced that they now have HBAs that are supported by VMware's architecture [6].

The first step is to drive out the upfront costs. Shared storage in the enterprise has typically had a high cost because it meant purchasing proprietary storage from a single vendor and implementing that shared storage on a new network infrastructure like Fiber Channel. While the costs of these systems and their infrastructures have come down, the purchase of something new will typically raise the cost of a project. Second, because the storage is from a single vendor there is often, (always?), a premium price associated with it.

For the SMB it would be more cost-effective if they could leverage the investment they already made in servers, storage and network. A way to accomplish this is to use separate storage software from the storage hardware. All storage systems, whether they are designed for small business or large enterprise, have essentially two components; hardware and software. The software is the intelligence that manages sharing of the storage and other capabilities like replication or snapshots.

Another key aspect in making shared storage affordable for the SMB Virtual Server project is to similarly leverage the existing network infrastructure instead of implementing a new one. Most SMBs will have some form of IP network in place long before they begin a virtualized server project. As a result, a storage system that can be shared with the existing IP networks would be able to further keep storage costs down even if the storage network can use its own switch or networking gear to ensure high performance of the storage traffic.

There are two storage protocols that leverage an existing IP infrastructure. The first is the file sharing protocol NFS, which is commonly found in NAS environments (file-level storage). NFS, while supported by some server virtualization platforms, is not supported by all. If it were to be used, for most SMBs, the purchase of a specific NAS based appliance would be required. For hosting virtual machine images a fairly powerful, and even more expensive, NAS is also required.

The second protocol that leverages the existing IP network and is used for block-level storage (i.e. SAN) is iSCSI. iSCSI encapsulates standard SCSI commands and sends them across an IP network. This means that almost anything that can be done with a local SCSI hard disk can also be done shared, across a network with iSCSI, including clustering and booting from the environment. For the small-to-medium-sized business this may be the ideal situation. iSCSI allows the SMB to leverage the existing IP network that they already have.

Once the storage platform is built, the SMB has to be confident that they can implement it and operate the environment. Since this iSCSI environment is built on a networking protocol that they are likely to be very familiar with, it is as simple as continuing to manage that network. Other than centralizing the storage and making it shareable, nothing new has been added. The SMB simply has another component on their existing IP network [7].

V. HOW STORAGE VIRTUALIZATION CAN ENHANCE SERVER VIRTUALIZATION FOR SMB's

Server virtualization creates a dynamic environment where large numbers of virtual machines are applied to applications. Storage virtualization makes it very easy to allocate capacity to these servers and then reallocate capacity as server's needs change.

A. Easy Provisioning of Volumes to Virtual Machines

In general, the value of virtualization grows with the number of servers being managed. Typically, therefore, virtualized environments support tens or even hundreds of logical machines on a very few, large physical boxes. Therefore, the ability to virtualize the storage to these virtual servers becomes very attractive. If a user has hundreds of virtual machines, where each server requires about ten volumes, the number of volumes required would be in the thousands. Having a SAN-based Volume Manager (SVM) on the back-end of the server virtualization, allows users to quickly and efficiently create volumes for each of the virtual machines. It eliminates the need to deal with LUN management at the array level. Due to the fact that each volume is an independent volume, those volumes can be mirrored, replicated be the source of snapshots, and even mounted into standard servers without the VMware operating system if needed.

Storage virtualization provides the flexibility to rapidly allocate capacity, and the ability to allocate thousands of volumes – as needed. Due to the fact that in a virtual environment it is very easy to add and remove virtual machines and applications, the environment becomes very dynamic. Virtual machines are created, used and then reallocated or removed. This provides tremendous flexibility. In this highly dynamic environment, it is also important that the storage be provisioned and reallocated after use with the same simplicity as the virtual servers. Storage virtualization enables this flexibility.

B. Test environments

Test Environments are one of the killer applications for virtual servers. A storage virtualization technology that supports low-capacity, point-in-time snapshots can increase this advantage by enabling the rapid creation of multiple snapshot copies of production volumes and their assignment to virtual test environments. Additionally, snapshots can reduce data preparation time before each test. So, testers can be assigned real ―live‖ data within seconds and then take snapshots of the data throughout the testing process. Should a multi-stage test fail at say, stage 13, the tester could go back to the snapshot taken at the beginning of the stage and run the test again, eliminating the need to repeat the 12 previous tests. Additionally, since the real failure may have occurred earlier in the testing process, the user could go back to previous snapshots taken at each stage and view the data to determine the root cause of the failure. All of these features significantly reduce the time needed for testing and increase the productivity of the testing team. Bringing a product to market quicker or isolating a software bug quicker can improve the profitability of a company.

C. Enhanced Backup

Having a virtual environment with hundreds of virtual machines can create a complicated, expensive backup proposition. Snapshot functionality obviates the need to install backup agents on every virtual machine. The backup can be done by creating snapshot copies for every virtual server and then assigning the copies to a virtual machine with the dedicated role of backup server. In this manner, the backup server is the only virtual machine that needs the backup software. When dealing with hundreds of virtual servers, this can reduce the cost of backup licenses considerably.

With capacity growing at exponential rates and processing hours becoming more and more important, backup windows are becoming non-existent. Simply stated, there is too much data to backup during off-hours. By comparison, snapshots can be taken at any time without taking the application offline. This creates a zero-window for backup. For many users, this solves the "shrinking backup" window problem.

Snapshots can add another significant benefit to the overall backup strategy. It is possible to keep point-in-time snapshots online for extended periods of time. If the data needs to be restored, the restoration can be done

in seconds from a point-in-time snapshot rather than in hours off tape, reducing the recovery time objective (RTO) from hours to seconds.

One issue in using snapshots is ensuring point-in-time consistency between volumes. This often requires that the application be suspended for a time while (for example) database buffers are flushed and all I/O completed. More complex solutions such as the use of consistency groups should be evaluated. Snapshots will not usually eliminate the requirement to take a full backup of the systems on a regular basis.

D. Consolidation of Servers at the Disaster Recovery Site

Today, remote mirroring and disaster recovery is a requirement even for very small companies due to regulation, corporate policies, or simply common sense. Large enterprises typically have the resources to spend on necessary communication lines, equipment, software, and training for disaster recovery. However, small and medium-sized businesses do not always have these resources, leaving the company exposed to regional disasters. This is another area where the combination of storage and server virtualization software can enable an affordable solution for disaster recovery.

Over the life of a remote mirror implementation, the most expensive component is the communication lines between the source and target locations. A solution that can remotely mirror using a snapshot-based technique, where only the differences between the snapshots are transmitted, avoids the need for very expensive communication lines between the locations. It is possible to use T1 or T3 lines that often cost hundreds of dollars per month, rather than higher bandwidth lines that can cost tens of thousands of dollars per month.

A remote mirror is an insurance policy to make systems available should a regional disaster render your primary site unavailable. However, there is no need to dedicate resources until a failure occurs. A virtual server gives users the flexibility to assign just very small amount of resources required to accomplish the mirror to the remote site. With snapshot-based remote replication, those resources are minimal because of the reduction of the amount of data being transmitted to the remote site. Should a failure occur at the primary site, server virtualization could then be used to assign more virtual machines to support the production workload. Virtualization of the remote resources allows them to be used for other purposes such as testing while they are in "standby" waiting for a failure, while guaranteeing that they will be instantly available when that primary site failure happens. This is true even if the primary site uses "physical" servers.

Thus the powerful combination of storage and server virtualization, therefore, allows users to build a disaster recovery site at a fraction of the cost. While large enterprises use expensive communication lines (e.g., OC-3, OC-12), high-end arrays in both the local and remote sites, and dedicated servers in the remote site, virtualization allows the use of inexpensive communication lines, inexpensive storage arrays at the remote sites and the use of virtual servers instead of physical servers at the remote location [8].

VI. CONCLUSION

Virtualization is a great opportunity to lower cost and raise productivity while reducing risk for businesses of any size and with budgets as low as zero. Many technologies promise important improvements for businesses but most create questionable value while incurring real cost. Virtualization brings real, measurable value while often costing nothing and often reducing spending immediately.

Virtualization technology is becoming more and more prominent in IT environments. Both server virtualization and storage virtualization offer unique benefits. However, it is not until the two technologies are combined that users truly recognize the full benefits of the solution. The ability to scale both server and storage resources as needed is a tremendous benefit. Additionally, the combination of the two technologies, create new opportunities for value that previously did not exist. Storage and server virtualization, when combined provide full resource virtualization to capitalize on these new benefits.

REFERENCES

[1] George Ou, "Introduction to server virtualization", May 2006

[2] Thomas Burger, " Intel Technology Journal: Special issue on virtualization technology", Volume 10, Issue 03, Oct 2008

[3] Richard Talber,"Using Virtualization to Improve Datacenter Efficiency", July 2009

[4] [4] Anup Pal, —Server Virtualization: Its Challenges and Benefits||, Oct 2010

[5] Brian Peterson, "SAN-based storage virtualization: Five benefits for your customers", Feb 2008

[6] Rick Cook, "Integrating storage and server virtualization", July 2006

[7] George Crump, "Why SMBs Should Combine Server Virtualization and Storage Virtualization", November 2009

[8] Nelson Nahum, " Combining Storage and Server Virtualization", Jan 2009

Network Payment Security
Using Intelligent Agent

Rupali H. Nikhare

Information Technology Engineering Department, Pillai's Institute Of Information Technology,
New Panvel, Navi Mumbai, Maharashtra, India

Abstract - At present, mechanisms used to ensure secure electronic transaction in E-commerce include SSL and SET. SSL provides a point-topoint data security between client and server, sending all the order information and user credit card information provided by online customer. However, SSL is usually so time consuming that it causes long time delay during data transmission and provides less security than SET at payment information. According to the different protocol, sever hardware, and the networked environments, SSL connections may slower 2-100 times than TCP/IP. On the other hand, the cost is the biggest shortcoming of SET. Many small businesses cannot afford it. This paper present the intelligent-agent based distributed system architecture and elaborate three kinds of task intelligent agents.

General Terms **:** Security

Keywords : E-commerce system, Task intelligent agents, Data Security, Adaptivity, Payment Security.

I. INTRODUCTION

In addition, the project we are now researching on has built up a virtual E-commerce system, but current SLL or SET software package cannot be seamlessly integrated with this system [3,4,5,6]. For all these reasons, we have been thinking of building up our own payment security scheme to ensure the payment security in our virtual E-commerce system.

Intelligent agents are a new paradigm for developing software applications. An agent is a computer system situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives. Autonomy is a difficult concept to pin down precisely, but we mean it simply in the sense that the system should be able to act without the direct intervention of humans (or other agents), and should have control over its own actions and internal state. It may be helpful to draw an analogy between the notion of autonomy with respect to agents and encapsulation with respect to objectoriented systems. An object encapsulates some state, and has some control over this state in that it can only be accessed or modified via the methods that the object provides. Agents encapsulate state in just the same way. However, we also think of agents as encapsulating *behavior*, in addition to state. An object does not encapsulate behavior: it has no control over the execution of methods – if an object *x* invokes a method *m* on an object *y*, then *y* has no control over whether *m* is executed or not – it just *is*. In this sense, object *y* is not autonomous, as it has no control over its own actions. In contrast, we think of an agent as having exactly this kind of control over what actions it performs. Because of this distinction, we do not think of agents as invoking methods (actions) on agents – rather, we tend to think of them requesting actions to be performed. The decision about whether to act upon the request lies with the recipient.

Of course, autonomous computer systems are not a new development. There are many examples of such systems in existence.

Examples include:

Any process control system, which must monitor a real-world environment and Perform actions to modify it as conditions change (typically in real time) – such systems range from the very simple (for example, thermostats) to the extremely complex (for example, nuclear reactor control systems).

An intelligent agent also has the ability of maintaining consistency of replicated, interdependent data across distributed sites. Conventional solutions use explicit data synchronization messages for consistency and replica management. An agent-based solution

allows the responsibility of consistency management to be assigned to mobile intelligent agents. An intelligent agent can also ensure secure distributed transactions,

that is, it can guarantee security and integrity of database transactions in a networked environment.

The use of intelligent agents in Ecommerce will not only contribute new solutions to some current problems, but also present a uniform platform and methodology for building such a distributed system.

II. LITERATURE SURVEY

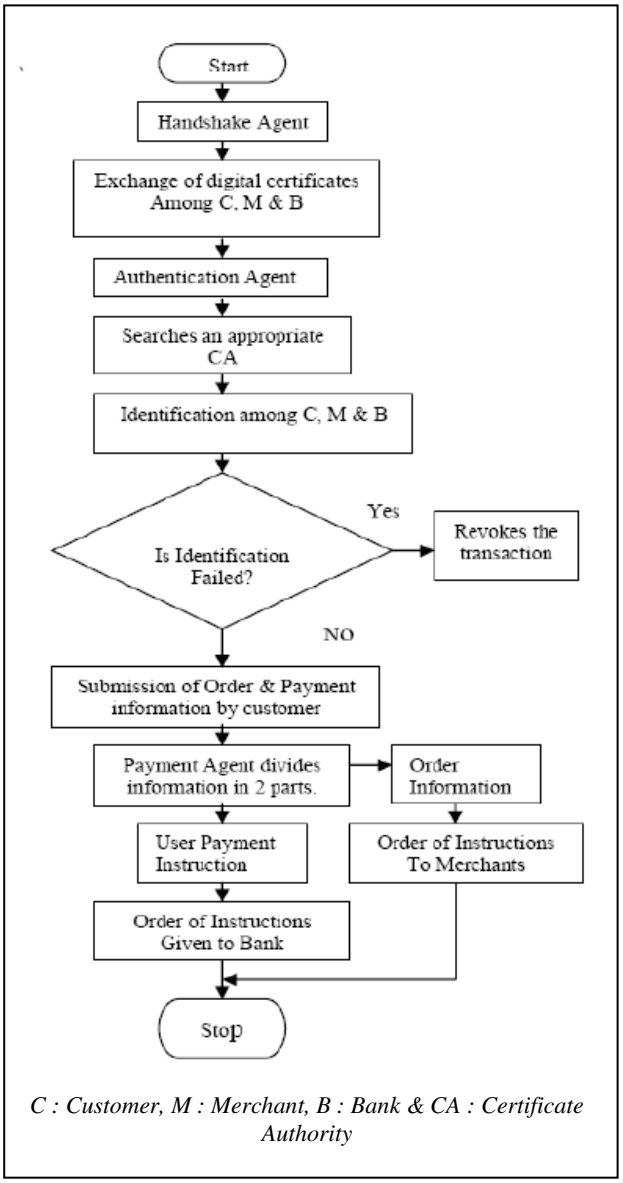
Intelligent agents are a new paradigm for developing software applications. More than this, agent-based computing has been hailed as 'the next significant breakthrough in software development' (Sargent, 1992), and 'the new revolution in software' (Ovum, 1994). An agent is a computer system situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives.

Intelligent agents are very suitable for building up distributed applications as they have the attributes of autonomy, adaptivity, collaborative behavior, inferential capability and mobility. An intelligent agent also has the ability of maintaining consistency of replicated, interdependent data across distributed sites. An agent-based solution allows the responsibility of consistency management to be assigned to mobile intelligent agents. An intelligent agent can also ensure secure distributed transactions, that is, it can guarantee security and integrity of database transactions in a networked environment. Intelligent agents have currently been used for workflow management, network management, air-traffic control, information retrieval management, etc.

III. FLOW DIAGRAM BASED ON INTELLIGENT AGENTS

The methodology is adopted to implement payment system architecture of the virtual E-commerce system using task agents [8]. The task agent in our payment security system is mainly composed of three kinds of intelligent agents, handshake agent, authentication agent and payment agent. The algorithms dedicated for encryption and decryption of payment information such as RSA algorithm. Fig.1 shows the whole flow diagram of payment security System.

The whole payment procedure of one electronic transaction is as follows. First, the client, the bank and the erchant exchange the electronic certificates through handshake agent. Second, the client, merchant and bank identify each other through authentication agent. If any of the three identifications fails, then revokes the transaction, else goes onto the next step. Third, the client submits his/her order and some additional information, then the payment agent will divide the information into two parts, one is the order information and the other is the user Payment Instruction. The payment agent will then automatically communicate with the payment agent in the bank server with the user Payment Instruction, and the payment agent in merchant server with the order information. After authenticating the user credit card, the bank server will do corresponding computing task and send "ok" information to the merchant server if it succeeds. Thus the merchant can confirm the user order and prepare for the commodity delivery.



C : Customer, M : Merchant, B : Bank & CA : Certificate Authority

Figure 1. Flow diagram of payment security system

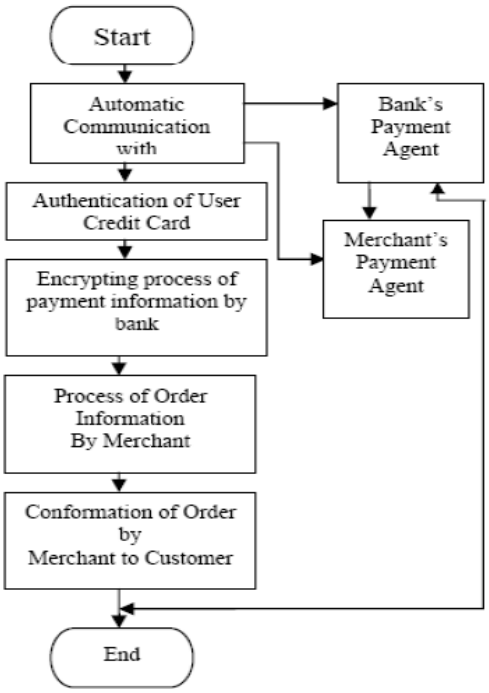


Figure 2. Transaction between bank and merchant

IV. HANDSHAKE AGENT

One of the task agent in our payment security system is handshake agent, which is designed to implement the handshake protocol in SSL among client, bank, and merchant. The handshake agent initiates some parameters with conducting information given by interface agent .

This includes four fields:

i) Encryption_algorithm_id: field is used to indicate the selected encryption algorithm.

ii) Public_key field : transmits the key the receiver used to encrypt its digital certificate serial number.

iii) Authentication _request field: is used to indicate whether the authentication is needed. All these fields in this structure are much smaller than the original structure. As a

result, the ServerHello structure is accordingly changed into StrResponse structure:

iv) Struct{

SessionID session_id;

EncyptedCertificateSN

certificate_serial_number;

} StrResponse

All these will reduce the spending of the

handshake, and also benefit for session recovery

V. AUTHENTICATION AGENT

The second step of an electronic transaction is to authenticate the digital certificates of all the participators. After the handshake procedure succeeds, the handshake agent will communicate with the authentication agent to identify the legitimacy of a participator.

The digital certificate structure used in authentication agent is:

Struct {

Version current_version;

SerialNumber

certificate_serial_number;

AlgorithmIdentifier algorithm_id;

Issuer CA_name;

IssuerUniqueIdentifier CA_id;

Subject user_name;

SubjectUniqueIdentifier user_id;

PublicKeyInfor

public_key_information;

PeriodofValidity deadline;

}StrCertificate

VI. PAYMENT AGENT

Payment gateway process is used in this network payment security system. Payment gateway protects credit cards details encrypting sensitive information, such as credit card numbers, to ensure that information passes securely between the customer and the merchant and also between merchant and payment processor. AES(Advanced Encryption Standard) algorithm is used for encryption and decryption of payment information.

Figure 2 and 3 exhibit encrypting process of payment information and process of order information.

C: Customer B: Bank

M: Merchant

PI: Payment Information

OI: Order Information

H: Hash Function

E: Encryption

KX,Y: The key share by X and Y

EKX,Y[M]: use the KX,Y encrypt the M

_: Connection

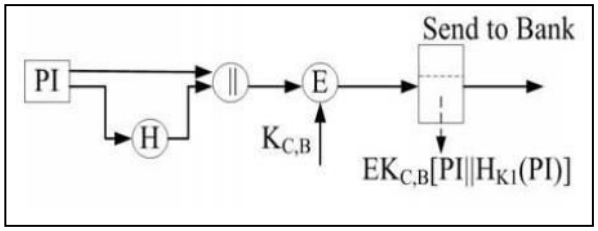


Figure 3. Encrypting process of payment information

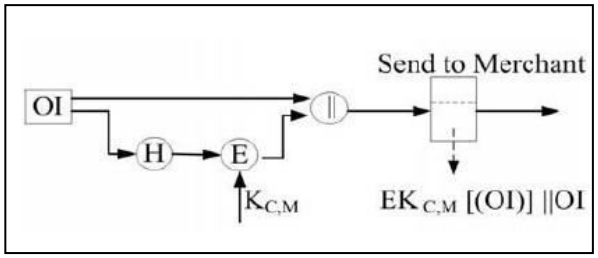


Figure 4. Process of Order Information

These two processes are implemented in a special task agent, payment agent. Fig. 4 shows the payment procedure of the payment agent.

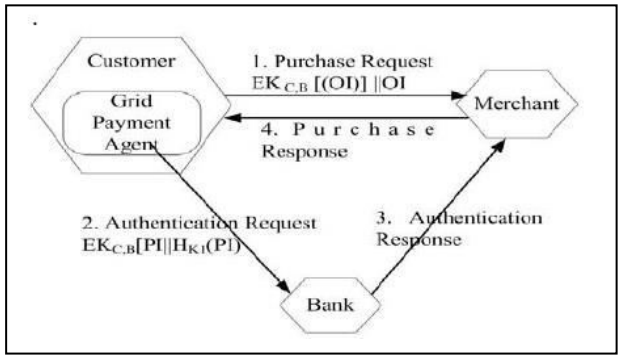


Figure 5. Payment procedure of the payment agent

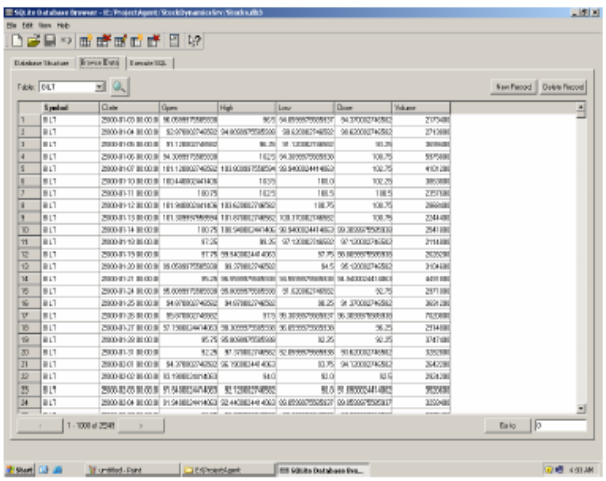


Figure 6. Database

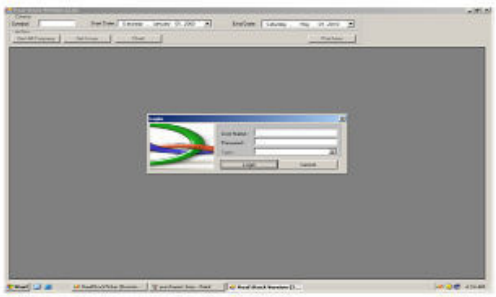
VII.SIMULATION RESULT

In this paper , we have shown an example of Share Market System.

7.1 Database

Database is implemented in SQL. It has seven fields such as Symbol, Date, Open Stock, High, Low, Close Stock and Volume. Symbol field indicates the name of companies. Open stock indicates the opening of the stock. Close stock indicates closing of the stock. Volume indicates the number of shares available for the sale.

7.2 Login Window



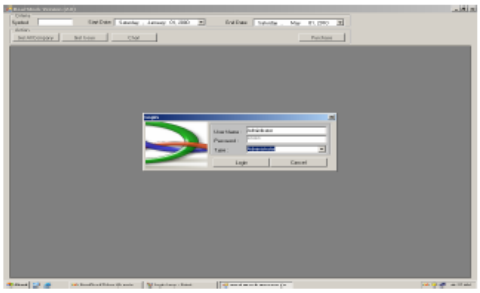
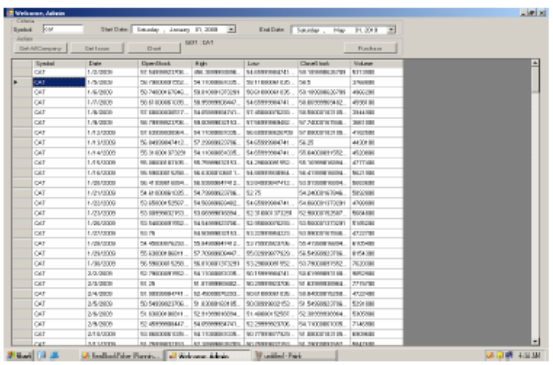


Figure 7. Login window

7.3 Result Window

Figure7. shows the result window which indicates records of CAT company for last 10 years and price chart for CAT company.



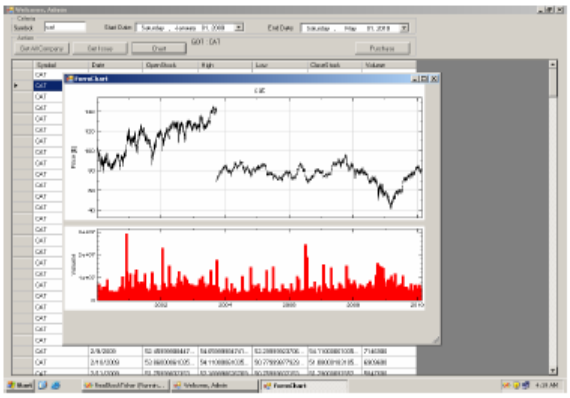
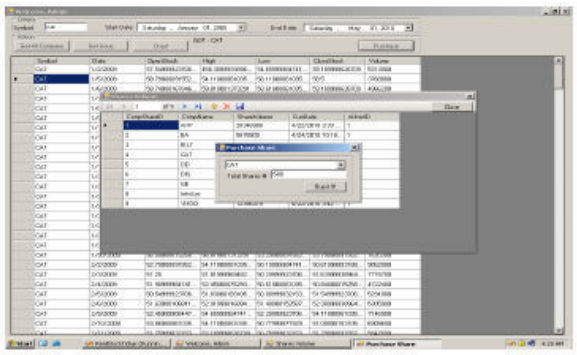


Figure 8. Result window



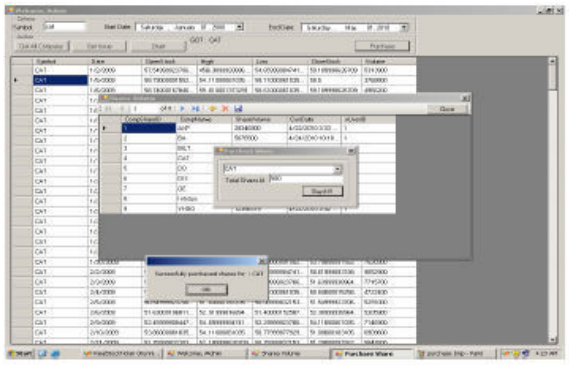


Figure 9. Successful Purchase For CAT Company

VIII. CONCLUSION

This paper present a scheme for the payment system, which is based on intelligent agents. The payment security scheme implemented in this paper is applying new methods to study and solve old problems. The application of intelligent agents in payment security can lead to a luciferous prospect.

REFERENCE :

[1] Eric Rescorla. SSL&TLS, 1st ed. Beijing: China Electronic Press, 2002, pp.202-205.

[2] Bruce Schineier. The truth in network information security, 1st ed. Beijing: Machine Press, 2001, pp. 216-217

[3] Jidi Zhao, Huizhang Shen. “Predictive data mining on web-based e-commerce store”, Proc. of IACIS’ 2002, Florida, U.S.A., 2002.

[4] Huizhang Shen, Daiping Hu, Huanchen Wang, A Group Decision Support System of Distributed Macroeconomics, Proceedings of International Conference on Management Science & Engineering, Moscow, HIT Press, 2001

[5] Huizhang Shen, Huanchen Wang, Group Decision Support System Based on Internet, Issues in Information Systems, Florida, U.S.A., 2002.

[6] Huizhang Shen, Huanchen Wang, Research on Distributed Model Management and Online Modeling, Proc. Of ICMSE, VOLS I AND II: 54-58˘2002

[7] Jennings, N.R. and Wooldridge, M. “Applications of Intelligent Agents,” in Agent Technology: Foundations, Applications, and Markets, N. Jennings and M.J. Wooldridge, Eds. Springer-Verlag, 1998, pp. 3-28.

[8] Hui-Zhang Shen, Ji-di Zhao, Application of Intelligent Agent in Network Payment Security, Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks (LCN’04)

◈◈◈

Scalable Design of Service Discovery Mechanism For Ad-hoc Network Using Wireless Mesh Network

Faiyaz Ahmad & Saba Khalid

Information Technology Department, Computer Science and Engineering Department
Integral University, Lucknow

Abstract - Wireless Mesh Network is an emerging technology that allow users to access information and services electronically using service discovery protocol. The seamless connectivity and mobility feature of WMN motivated us in the design of efficient and scalable service discovery scheme that assures certain level of quality of service. The proposed model uses routing clients which communicate with Service Caches to register for services. The gateway nodes discovers quality services by using backbone based distributed directory structure. The proposed model is scalable and reduces discovery overhead, duplicate information dissemination, and energy consumption.

Keywords- *Wireless Mesh Network, Service Discovery, Service Directories, Service caches, OFLSR,routing deamon*

I. INTRODUCTION

Wireless Mesh Networks (WMNs) provide flexibility in terms of mobility i.e. Mesh clients can be stationary or mobile and can form a client mesh network among themselves and with mesh routers. WMNs make use of multiple radios and multiple channels per radio for increased capacity, higher throughput and low interference.[1]. Service discovery is acclaimed as a crucial challenge in WMN [2], [3], [4]. Service discovery is the ability to discover and form an ad-hoc network without explicit user direction. It facilitates devices and services to properly discover, configure, and communicate with each other. Service discovery minimizes administrative overhead and increases usability [4]. So, Service discovery can be defined as a process that allows networked entities to broadcast their services, inquire about services provided by other clients, selecting the desired service and invoking it. In the literature, many service discovery schemes have been provided. Current approaches for service discovery uses different directory based architectures for flooding information into the network. In literature there are two architecture approaches to perform service discovery – (i) directory based and (ii) non directory based approaches. In later approach broadcasting or multicasting is performed for flooding messages in the network.. The proposed scheme uses the mesh backbone approach and selects a set of nodes on the basis of stability constraints to coordinate the network. The nodes which have minimal mobility are elected as mesh backbone and gateways or routers. These nodes function as service caches. This paper proposes the design of an efficient and scalable service discovery scheme that optimizes discovery overhead by integrating discovery information in routing demon. We have also implemented the functionality of OFLSR routing protocol in the network layer this reduces the routing update overhead by using different exchange period for different entries in routing table. The advantage of this scheme is that it reduces message overhead, battery power consumption and maintenance messages for SC. The remainder of the paper is organized as follows: Section II presents related work and existing techniques for service discovery. Section III provides detail description of proposed model. Section IV discusses working of proposed architecture Section V discusses and evaluates scalibity issues. Finally in section VI paper concludes with future work.

II. RELATED WORK

On the protocol level there are a lot of well-known protocols for service discovery like the Service Location Protocol version 2 (SLPv2) [5], Simple Service Discovery Protocol as part of Universal Plug and Play (UPnP) [6] or the DNS based Service Discovery (DNS-SD) [7]. However, these protocols were originally

designed for wired LANs or small single-hop ad-hoc network. mSLP [8] is one of many existing modifications to SLP where SLP Directory Agents form a mesh structure and exchange service registration states. However, this approach does not scale well in WMNs, because service registration states must be replicated between all servers.

A light-weight service discovery (LWSD) protocol for ad hoc networks was proposed by Mallah and Quintero [9]. This protocol deploys judiciously elected (stable) nodes in the environment for the service discovery to take place. Artail et. al describes a distributed service discovery model (DSDM) for MANETs [10]. This model is similar to the approach in [9]. Virtual backbone nodes act as service directories nodes. This model handles network partitioning. The intermediate nodes between the service provider and service requestor cache the service information. This model minimizes the use of packet flooding and broadcasting for service advertisement and discovery. This model achieves reasonable system response time and network traffic. This model uses proactive routing protocol to update the routing table for frequent disconnections and arrivals. This model does not consider high node mobility on system performance.
The proposed model extends the work of [9] [11] [12] [13] to achieve quality of service(QoS) based service discovery on integrated directory-base architecture.

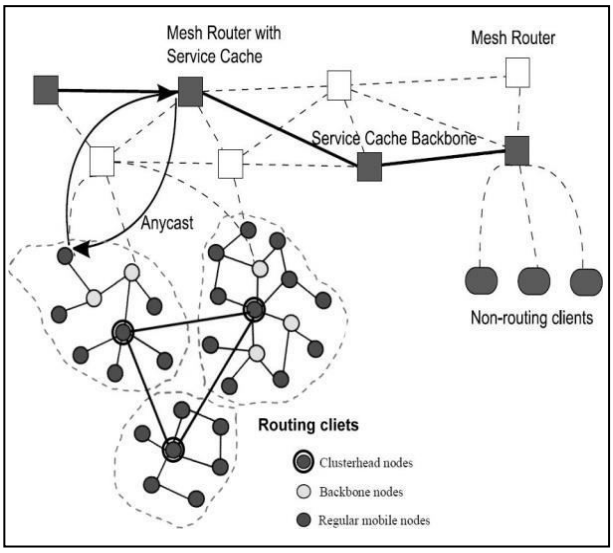


Figure 1. Back bone based directory architecture of WMN

III. PROPOSED WORK

The proposed system framework for QoS discovery in WMN is shown in figure 2. First, bootstrapping process is performed by node on application layer which helps in neighbor nodes and services identification. Then mesh network is organized using integrated directory-based architecture, backbone-based and cluster-based. After mesh backbone routers and clients discovery we have implemented Optimised Fisheye Link State Routing Protocol(OFLSR) which combines two routing protocols OLSR and FSR. It(OFLSR) divides the network into different scope levels for service advertisement and discovery. Implementation of this protocol significantly reduces duplicate packet retransmission. The service discovery component comprises of functional description and assigns QoS values dynamically. The service descriptions are registered by service providers to their clusters or zones or scopes. Then services are discovered within scopes either locally or globally. The network is reconfigured when either new mesh nodes join clusters or leave the network. Below are listed various components of proposed service discovery scheme in WMN using OFLSR protocol.

A. Service Discovery Scheme Components

1. Clients: are the entities that either provides services to other entities or expect to discover available services in an unknown environment with the help of node advertisement and neighborhood discovery.

a) Node Advertisement for Neighborhood Discovery: Initially in WMN there are no backbone nodes(BB) or clusterheads(CH). Every node in the network perform lightweight advertisement of HELLO messages to discover neighbor nodes. This message usually contains the following information fields

- Source∟ ⊥∟Net_id of node, node_id, BB, CH, packet_type)
- ∟List of next hop neighbors (Network_id, node_id, BB,CH, latest Time stamp)
  ∟Control information (Time To Live)

Initially the value of backbone (BB) nodes and cluster head (CH) nodes will be zero.

2) Directory

Directories are entities that are responsible for caching advertisements from available service provider nodes(SP) and carries out lookup to cater to discovery requests from clients. Explicit directory agents are used by in approaches. The proposed backbone based directory module consists of three types of nodes- Service Directories (SD) nodes, Service Provider (SP) nodes and Regular non routing nodes. SD nodes are BB nodes and CH nodes and they act as service caches(SC).

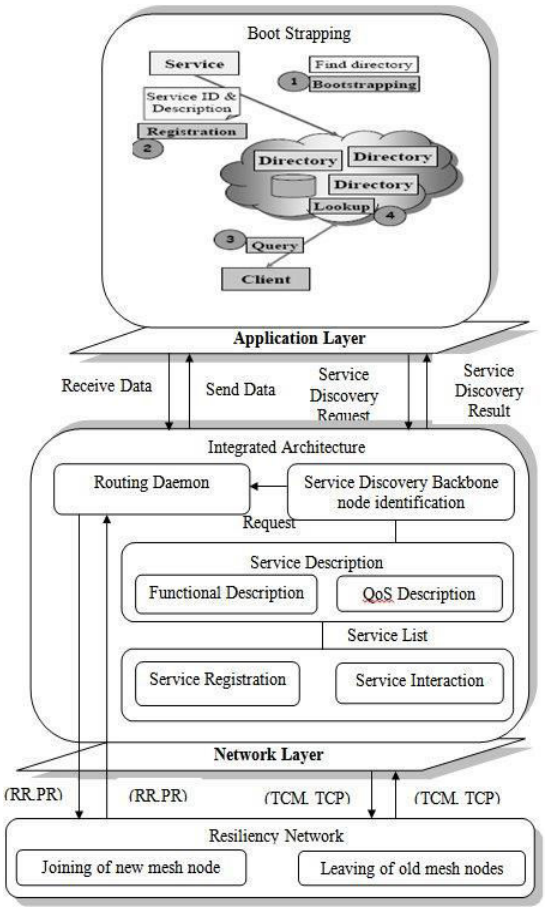


Figure 2. Proposed Service Discovery Scheme design for WMN using OFLSR Protocol in Network Layer.

a) Backbone Nodes Identification

First BB nodes are randomly identified on the basis of stability constraints namely as Normalized Link Failure Frequency (NLFF) metric[9]. NLFF reflects the rapidity in which the neighborhood changes in comparison with its degree.

Then clustering technique or scope formation is applied according to Fisheye State routing to form integrated directory-based structure. This structure can be used to discover services among nodes.

b) CH identification or Proactive zone head formation

Only the BB nodes are eligible to be CH. CH node is a subset of BB node and form service caches. BB node sends chello message to the immediate minimal neighbor nodes with TTL field set to some value say x, this message will reach to local neighbor nodes only, they send this message to their local neighbors. These all nodes will then form a cluster.

3) Service

It abstracts a set of functionalities offered by a networked entity. Service description comprises of information like service name, service capabilities, non-functional attributes (QoS parameters). QoS parameters are reliability, security, response time, latency, throughput, correctness and availability. The Ontology Web Language for Services (OWL-S) [10] is used for functional description of services and WS-QoSOnto [14] is used for QoS description.

IV. WORKING OF SERVICE DISCOVERY SCHEME

A) Bootstrapping

It is the first step in service discovery and some a priori information is required. This process uses unicast and multicast communication techniques to declare its existence in the network and to avail the services of the WMN.

B) Service Registration

The mesh clients are service requestors (SRs). SR nodes requests for services from SD which are BB nodes and CH nodes. Each scope consists of one CH node, one or more BB nodes and other non routing nodes. Service Providers (SP) nodes register their services to one of nearest SD nodes in its scope. This SD node, performs lookup using service list, and is responsible to distribute the received services via unicast or multicast to other SD nodes in its zone. The service registration is done locally in its scope. Thus, this scheme limits the flooding of the service registration packets to local BB nodes by adopting OFLSR technique since a source only needs to know the approximate route towards the destination far away. In this way, all SD nodes in scope will maintain their list of services offered by their scope level members. The SP nodes must renew their registration periodically with any of SD node they can reach.

C) Service Discovery

The services can be discovered using routing deamon and QoS description shown in figure 2. When an application program installed at a mesh client wants to find an appropriate service, a unicast service discovery request is sent to the SD module in the local mesh router(CH). The mesh client is neither aware how many SD are in the network nor has any control over the decision to which SC the query is routed. OFLSR routes the unicast query to the best SD. Due to route switching it could happen that a client is directed to different SCs each time it queries. The decision of choosing an appropriate SC is made in the network layer. The SC verifies whether the requested service exists or not in the

service list. If this SD node finds the requested service description information, then it sends reply to the requesting node. This model optimizes the flooding of service request packets to reach only local SD nodes in a scope rather than all BB nodes in the WMN. If the requested service does not exist in the service list then the SC module requests the routing deamon to collaborate in order to find the requested service by generating the appropriate route request packet and by passing the application request to the integrated architecture that will generate the appropriate route discovery packet(RP). Routing deamon forwards this RP until on demand cache replication is performed in network, and CH sends service reply. A CH has two main tasks: Handling service queries and registrations from mesh clients and communication with other CHs in the backbone.

1) Registrations: When a CH receives a unicast service registration from a client, the CH stores the service record with a timestamp in its cache. If the timestamp has expires, the client has to re-register or the corresponding service record is deleted from the integrated architecture. Service records are not replicated to other CHs after a new registration until they are requested. If the CH receives new services from other CHs, it sends the new services to the client in route reply packet.

The advantage of this scheme is that services are only replicated on-demand, i.e. the information is pushed to all CH if another CH queries for the service. This prevents useless replications of services which are rarely or never used, but it increases request delay. It also increases packet efficiency, because multiple service records are aggregated into one.

D) Resiliency network

When new node approaches the network a topology control message (TCM) which contains topology information necessary to build routing table is sent to the routing daemon. The OFLSR protocol present in our design limits the flooding of the TCM by adopting FSR technique since a source only needs to know the approximate route towards the destination far away. When old mesh client leaves the network environment no TCM are flooded in WMN, since OFLSR does not triggers TC messages on broken links or link failure, thus reducing routing update traffic. Hence network is resilient on new nodes entry and old nodes exit.

V. PERFORMANCE EVALUATION

We performed simulations in ns2 to ensure that the proposed service discovery scheme design works and is feasible for wireless mesh networks. The first step was simulation of a WMN to demonstrate that networks with different radio technologies can communicate with each other. In the next step, developed mechanism named as Integrated Discovery architecture was tested. This scheme is tested under various mobility conditions and different node topologies. The efficiency of this scheme can be evaluated with network load (in terms of number of packets),average time delay (between the time any successful request is sent from a client and the time a corresponding reply is received by the same client) and battery power consumption. The proposed scheme partitions the network into different scope levels. The scheme uses Optimized Fisheye Link State Routing protocol. In OFLSR, the reduction of routing update overhead is obtained by using different exchange periods for different entries in routing table[14]. More precisely, service request corresponding to nodes within the smaller scope are propagated to the neighbors with the highest frequency. OFLSR limits the flooding of the TC message hence is resilient to increasing traffic load, since it does not repair broken links hence provide better throughput. Its exhibits a much better scalability of traffic load compared to the scheme proposed by [11][12][13].

VI. CONCLUSION AND FUTURE WORK

In this paper we provided a quality based service discovery scheme for wireless mesh networks. The proposed scheme implements OFLSR protocol which provides a better performance in terms of data packet delivery ratio throughput, packet latency and routing overhead, under different traffic and mobility instances. This scheme achieves reduced network load, reasonable mean time delay to the requests initiated by the clients, great average hit ratio of successful attempts, reduced battery power consumption. The scheme design uses service caches which optimizes the flooding of packets during service registration and discovery. Moreover, the cost (in time, control packets) to construct and maintain the clusters is almost negligible, Further, services are shared among SD nodes to reduce the overhead. The services are registered locally in zones and they are discovered with respect to QoS criteria across clusters. This model needs to address network maintenance. Moreover, our protocol outperforms existing protocols by reducing the network overhead. This makes our protocol efficient and scalable in WMNs.

Future direction of work will be to further study our scheme from other perspectives, such as reduced service response time and increased service availability for popular services.

REFERENCE :

[1] Shahibzada Ali Mehmood,Shahbaz Khan,Sohaib Khan,,Hamed Al Raweshidy"A comparison of MANETs and WMNs: commercial feasibility of

community wireless networks and MANETs"in AcessNets '06 Proceedings of the 1st international conference on Access networks,ACM

[2] F. Casati, S. Ilnicki, L.J. Jin, and Hewlett-Packard Laboratories. Adaptive and Dynamic Service Composition in EFlow. Springer, 2000.

[3] H. Chen, A. Joshi, and T. Finin. Dynamic Service Discovery for

[4] Mobile Computing: Intelligent Agents Meet Jini in the Aether. Cluste Computing, 4(4):343–354, 2001.

[5] B. Pascoe. Salutation-Lite. The Salutation Consortium June, 6, 1999.

[6] E. Guttman, C. Perkins, J. Veizades, and M. Day, "Service Location. Protocol, Version 2," RFC 2608, June 1999.

[7] "Universal Plug and Play (UPnP)." [Online]. Available: http://www. upnp.org/

[8] S. Cheshire and M. Krochmal, "DNS-Based Service Discovery," IETF Internet Draft, August 2006, work in progress.

[9] W. Zhao and E. Guttman, "mSLP - Mesh Enhanced Service Location Protocol," IETF Experimental RFC 3528.

[10] R. A. Mallah and A. Quintero, "A Light-Weight Service Discovery Protocol for Ad Hoc Networks", Journal of Computer Science, Science Publications, Vol. 5, No. 4, pp. 330 – 337, 2009.

[11] H. Artail, K.W. Mershad, and H. Hamze, "DSDM: A Distributed Service Discovery Model for Manets", IEEE Trans. Parallel and Distributed Systems, Vol. 19, No. 9, pp. 1224-1236, Sept. 2008.

[12] Martin Krebs, Karl-Heinz Krempels" Optimistic On-demand Cache Replication for Service Discovery in Wireless Mesh Networks" in Proc. Of IEEE, 2009

[13] R. Deepa, S. Swamynathan" A Service Discovery Model for Mobile Ad hoc Networks in Proc. Of IEEE , International Conference on Recent Trends in Information, Telecommunication and Computing,2010.

[14] Kaouther Abrougui, Azzedine Boukerche "A Mesh Hybrid Adaptive Service Discovery Protocol (MesHASeDiP): Protocol Design and Proof of Correctness" in Proc of IEEE,2008

[15] Jiwei Chen, Yeng-Zhong Lee, Daniela Maniezzo, Mario Gerla" Performance Comparison of AODV and OFLSR inWireless Mesh Networks"in Proc of IEEE

◈◈◈

Cloud Computing

Kalidas Nalla, Mohit Saxena & Mannepalli Kailash
Dept. of Computer Science and Engineering, SRM University, Chennai, India

Abstract - Cloud computing is clearly one of today's most enticing technology areas due, at least in part, to its cost efficiency and flexibility. Cloud computing can and does mean different things to different people. However, despite the surge in activity and interest, there are significant, persistent concerns about cloud computing that are impeding momentum and will eventually compromise the vision of cloud computing as a new IT procurement model. In this paper, we characterize the problems and their impact on adoption. In addition, and equally importantly, we describe how the combination of existing research thrusts has the potential to alleviate many of the concerns impeding adoption. This publication also provides an overview of the security and privacy challenges pertinent to cloud computing and points out considerations organizations should take when outsourcing data, applications, and infrastructure to a cloud environment which would only make the cloud reliable and gain prominent interest amongst the cloud prospective players and users. Life certainly would be better and beneficial inside the cloud and that is the reason this is our topic of discussion.

Keywords: *Cloud Service Provider (CSP), Cloud security,European Union(EU), Data Protection Act (DPA)*

I. INTRODUCTION

Cloud computing is certainly the most widely discussed area of evolving technology in the recent era. Though it has been a headache to the security managers of companies, the very aspect that the users pay to what they use rather than pay for everything is what has invoked a plethora of interest and enthusiasm among the technologists around the world. The field of cloud is a relatively new and is still in the developmental stages, which is why it has proved a lot of interest which in turn involves a huge number of IT technocrats always trying to improve it's services.

The definition of cloud is not fixed and it varies by the way an individual sees it. The definition normally referred to with cloud computing is that Cloud computing is Internet based development and use of computer technology. In concept, it is a paradigm shift whereby details are abstracted from the users who no longer need knowledge of, expertise in, or control over the technology infrastructure ,in the cloud that supports them. It typically involves the provision of dynamically scalable and often virtualized resources as a service over the Internet. According to Gartner, a well known research and advisory firm it is defined as "a style of computing where scalable and elastic IT-enabled capabilities are delivered as a service to external customers using Internet technologies."

The purpose of this paper Cloud Security is to provide needed context to assistorganizations in making educated risk management decisions regarding their cloud adoption strategies.

CLOUD SECURITY IS

- The response to a familiar set of security challenges that manifest differently in the cloud.
- A set of policies, technologies, and controls designed to protect data and infrastructure from malicious users and enable regulatory compliance.
- Layered technologies that create a durable security net or grid. Security is more effective when layered at each level of the stack and integrated into a common management framework.

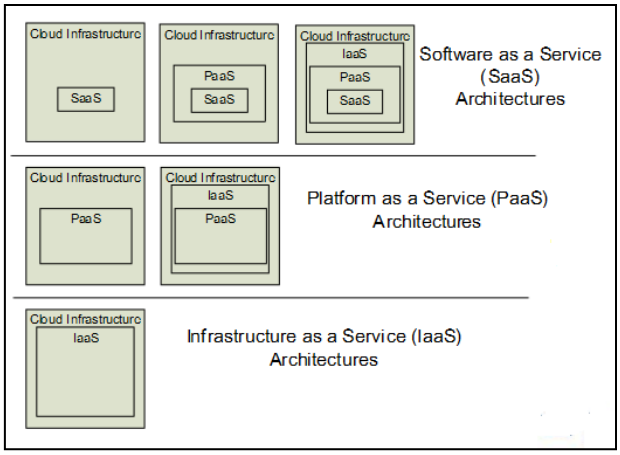
Cloud provides a virtualized environment for the user in the sense that users select the services needed and puts them into the cloud and has access to the data at anytime, anywhere, any using devices.Cloud computing can be divided into three categories for easy understanding.

Infrastructure as a service (IaaS), which provides flexible ways to create, use and manage virtual machines (VMs).Delivers computer infrastructure as a utility service, typically in a virtualized environment. Provides enormous potential for extensibility and scale

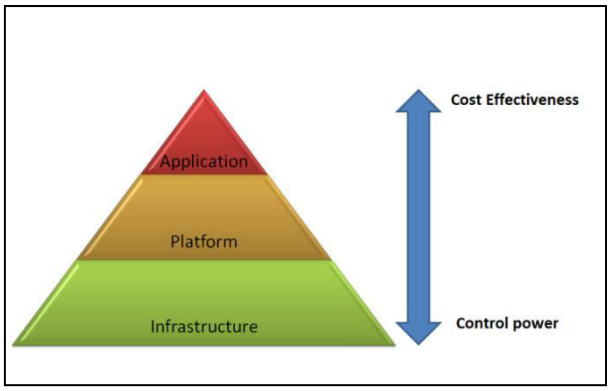
Platform as a service (PaaS), focuses on providing the higher-level capabilities. Delivers a platform or solution stack on a cloud infrastructure sits on top of the IaaS architecture and integrates with development and middleware capabilities as well as database, messaging, and queuing functions.

Software as a service (SaaS),are the applications that provide business value for users. Delivers applications

over the Internet or intranet via a cloud infrastructure built on underlying IaaS and PaaS layers.



The pyramid gives an ample idea of the inverse relationship between the cost efficiency and control. As we move up the pyramid we have limited access to our data, on the converse moving down gives us the freedom to deploy our device and access it. It all depends on the way the enterprise wants to avail the services of the cloud.



II. SECURITY ISSUES

In the last few years, cloud computing has grown from being a promising business concept to one of the fastest growing segments of the IT industry. Now, recession-hit companies are increasingly realizing that simply by tapping into the cloud they can gain fast access to the ingenuity of the business applications or drastically boost their infrastructure resources, all at negligible cost. But as more and more information on individuals and companies is placed in the cloud, concerns are beginning to grow about just how safe an environment it is.

2.1) **Understand the risks of cloud computing**

Companies need to be vigilant, for instance about how passwords are assigned, protected and changed. Cloud service providers typically work with number of third parties, who any have access to the data stored within the cloud.

Remediation's:

Customers are advised to gain information about those companies which could potentially access their data, and ensure that strong policies are entrusted upon them to ensure that the data is not misused.

2.2) **How cloud hosting players have seen security**

Companies need to know, for instance, whether a software change might actually alter its security settings. IBM, Cisco, SAP, EMC and several other leading technology companies announced that they had created an 'Open Cloud Manifesto' calling for more consistent security and monitoring of cloud services. But the very fact that neither Amazon.com, Google nor Salesforce.com agreed to take part suggests that broad industry consensus may be some way off. Microsoft also abstained from the conference, charging that IBM was forcing its agenda.

Remediation's:

Austerenorms should be taken by the company to ensure the trust of the customer. Few of the points to consider by the Cloud Service Providers (CSP) are:

- Inquire about exception monitoring systems.
- Be vigilant around updates and make sure that staffs don't suddenly gain access privileges they're not supposed to.
- Be careful to develop good policies around passwords; how they are created, protected and changed.
- Look into availability guarantees and penalties.
- See if the encryption key is accessible by any other third party.

2.3)**Local law and jurisdiction where data is held**

Possibly even more pressing an issue than standards in this new frontier is the emerging question of jurisdiction. Data that might be secure in one country may not be secure in another. Malicious users like terrorists may bank upon this loose point to plot terror attacks.

The DPA does not prohibit the over seas transfer of personal data, but it does require that it is protected adequately wherever it is located and who ever is processing it. Clearly, this raises compliance issues that organizations using internet-based computing need to address. EU favours very strict protection of privacy

Remediation's:

• Overseas data transfer must be given permission only to CSPs having agreed to be under the vigilance of Government. In America laws such as the US Patriot Act invest government and other agencies with virtually limitless powers to access information including that belonging to companies.

2.4) **Malicious Insiders**

The threat of a malicious insider is well-known to most organizations. A cloud service provider may not reveal how it grants employees access to physical and virtual assets, how it monitors these employees, or how it analyzes and reports on policy compliance. To make matters worse, there is often little or no visibility into the hiring standards and practices for cloud employees. This kind of situation clearly creates an attractive opportunity for an adversary ranging from the hobbyist hacker, to organized crime or even a corporate espionage. The level of access granted could enable such an adversary to harvest confidential data or gain complete control over the cloud services with little or no risk of detection.

Remediation's:

- Enforce strict supply chain management and conduct a comprehensive supplier assessment.
- Specify human resource requirements as part of legal contracts.
- Require transparency into overall information security and management practices, as well as compliance reporting.
- Determine security breach notification processes.

2.5) **Account or Service Hijacking**

Attack methods such as phishing, fraud, and exploitation of software vulnerabilities still achieve results. Credentials and passwords are often reused, which amplifies the impact of such attacks. If an attacker gains access to your credentials, they can eavesdrop on your activities and transactions, manipulate data, return falsified information, and redirect your clients to illegitimate sites. Your account or service instances may become a new base for the attacker. From here, they may leverage the power of your reputation to launch subsequent attacks.

Remediation's:

- Prohibit the sharing of account credentials between users and services.
- Leverage strong two-factor authentication techniques where possible.
- Employ proactive monitoring to detect unauthorized activity.
- Understand cloud provider security policies and SLAs .

2.6) **Multitenancy and shared technology issues:**

Multitenancy refers to the ability of services to be offered to multiple tenants in away so that each tenant operates as logically isolated while, in fact, using physicallyshared resources. That means shared infrastructure CPU caches, graphics processing units (GPUs), disk partitions, memory, and other components that was never designed for strong compartmentalization. Even with a virtualization hypervisor to mediate the access between guest operating systems and physical resources, there is always a concern that attackers can gain unauthorized access and control of your underlying platform with software-only isolation mechanisms. Potential compromise of the hypervisor layer can in turn lead to a potential compromise of all the shared physical resources of the server that it controls, including memory and data as well as other virtual machines (VMs) on that server.

2.7)**Data loss or leakage:**

Protecting data can be a headache because of the number of ways it can be compromised. Confidential data like customer's bank account number, social security number should be protected from unauthorized users to prevent any kind of misuse. Data can also be maliciously deleted, altered, or unlinked from its larger context. Loss of data can damage the company's brand and reputation, affect customer and employee trust, and have regulatory compliance or competitive consequences.

Examples

Insufficient authentication, authorization, and audit (AAA) controls, inconsistent use of encryption and software keys, operational failures, risk of association, jurisdiction and political issues; data centre reliability; and disaster recovery.

Remediation's:

- Implement strong API access control.
- Encrypt and protect integrity of data in transit.
- Analyze data protection at both design and run time.
- Implement strong key generation, storage and management, and destruction practices.

- Contractually demand providers wipe persistent media before it is released into the pool.
- Contractually specify provider backup and retention strategies.

2.8) **Unknown risk:**

One of the dogmas of Cloud Computing is the reduction of hardware and software ownership and maintenance to allow companies to focus on their core business strengths. This has clear financial and operational benefits, which must be weighed carefully against the contradictory security concerns complicated by the fact that cloud deployments are driven by anticipated benefits, by groups who may lose track of the security ramifications.

Versions of software, code updates, security practices, vulnerability profiles, intrusion attempts, and security design, are all important factors for estimating your company's security posture. Information about who is sharing your infrastructure may be pertinent, in addition to network intrusion logs, redirection attempts and/or successes, and other logs. Security by obscurity may be low effort, but it can result in unknown exposures.

Remediation's:

- Disclosure of applicable logs and data.
- Partial/full disclosure of infrastructure details.
- Monitoring and alerting on necessary information.

III. SECURITY MEASURES

Security is the most important aspect when data is stored at an unknown location. Hence security should be integrated early with the cloud setup planning. Security considering the importance cannot be a one-step process.Your security profile in the cloud is defined by what your organization needs and the workloads you plan to move to the cloud.

One of the key elements for cloud security is to build in balanced controls instead of designing an environment based primarily on preventive controls. For example, the architecture is designed to dynamically adjust the user's access privilege as the level of risk changes, taking into account factors such as location and type of device. This new approach is both provocative to minimize occurrences and reactive to minimize damage if a breach occurs. Here are few of the aspects on cloud security to be considered foremost.

- Are your physical compute resources located on-premises or off-premises?
- What types of assets, resources, and information will be managed?
- Who manages them and how?
- Which controls are selected, and how are they integrated into the overall cloud architecture?
- What compliance issues do you face?

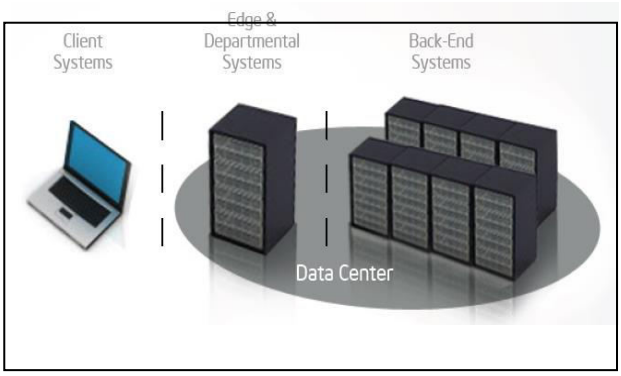
3.1) **Early Planning**

- Identify the business priorities for moving the specific workload(s) to the cloud. Security concerns can be weighed more effectively once the business context is defined for purpose of moving workloads to the cloud.
- Evaluate the sensitivity of the asset(s). This helps to understand the importance of the data or function. This evaluation can be made as a rough assessment or follow a specific valuation process.
- Map the security workload to the appropriate cloud delivery model and hosting models under consideration. Upon understanding the importance of the assets, evaluation the risks associated with various deployment models takes place.
- Determine whether the available services are capable of meeting the requirements for handling data, especially for compliance purposes. At this point, there arises a need to evaluate risk tolerance for the workload. If cloud service provider is known, a more detailed risk assessment can be conducted.
- Map the data flow, especially for public or hybrid cloud providers. For specific deployment options, data flow between your organization, the cloud services, and any customers needs to be known.

3.2) **Vulnerability Assessment of the Service Selected**

Cloud computing depends heavily on virtualization to realize operational savings and efficiencies, has elastic boundaries, and potentially pushes out the perimeter of the enterprise and security controls far beyond the data center.

Regardless of the cloud delivery model you choose, your best approach is to review the specific service architecture, and then layer technologies to develop a strong security net that protects data, applications and platform, and network at all levels. Because the model for your cloud services may be very different from other organizations and indeed may evolve and change over time it is recommended that, in addition to security software solutions and application features, you should strengthen your security net by protecting data and platform at the most basic level, the system hardware.



The above illustration shows how protection at the hardware level can enable security deeper in the data center. Compute resources complement your perimeter controls, enable more advanced security and compliance capabilities in existing solutions, and provide needed protection even below the hypervisor an area of emerging threat.

3.3) **Vulnerability Mitigation**

With protection at the hardware level, you can build trust andcompliance into your data center. This means you can:

- Provide the foundation for a more powerful layered security netof solutions and software features
- Put more granular controls closer to where your data is androbust platform services.
- Control where the VMs are distributed
- Protect confidential data and meet compliance requirements.It is highly recommended to prioritize your security investment through a risk assessment to determine the order and timing for building this level of trust and compliance into your data center in four areas.
- Encrypt to protect data.
- Establish a trusted foundation to secure the platform and the infrastructure.
- Build higher assurance into auditing to strengthen compliance.
- Establish and verify identities before you federate by controlling access to trusted clients from trusted systems.

3.4) **Data Protection using Encryption**

Encryption is an effective, well-established way to protect sensitive data because even if information is lost, it remains unusable. There are a number of ways to perform encryption, but typically it comes with a cost what is often referred to as a performance tax. Encryption is applied on:

1. **Data in motion**

- Data in flight over networks (Internet, e-commerce, automated teller machines, and so on)
- Data that uses protocols such as Hypertext Transfer Protocol Secure , FTP, and Secure Shell (SSH)

2. **Data in process**

- Transactional data in real time, such as encrypted fields, records, rows, or column data in a database.

3. **Data at rest**

- Files on computers, servers, and removable media
- Data stored using full disk encryption and application-level models.

3.5) **Start from the scratch-the platform**

Rootkit attacks are increasing. They are difficult to detect with traditional antivirus products and use various methods to remain undetected. Rootkit attacks infect system components such as hypervisors and operating systems, and the malware can operate in the background and spread throughout a cloud environment, causing increasing damage over time.

The best way to secure your platform is to enable a trusted foundationstarting with a root of trust at the platform level and extending the chain of trust through measured firmware, BIOS, and hypervisor virtualization layers. It enables a more secure platform for adding tenants and workloads. Essentially you build protection into your hardware to protect your software.The root of trust enables a trusted foundation within your cloud environment so you can:

- Specify trusted server pools. You can make decisions about how much to expose your data and workload based onwhether a trusted pool is established. The most sensitive workloads should always use a trusted pool.
- Respond quickly to attack and minimize damage. Detectattacks more quickly, contain the spread of malware, and reduce the need to rebuild hypervisors if a compromise is detected.

3.6) **Inter-cloud Trust**

With evolution of the cloud, the vision of federated cloudsacross which communications, data, and services can move easily within and across several cloud infrastructuresadds another layer of complexity to the

security.Solutions that extend trust across federated clouds via secure gateways between the service provider and the service consumer with policy enforcement for centrally defined policies.

Software solution can be used to control the entire life cycle of secure access for the enterprise connecting to cloud. It records the user activity against the systems, and the metrics can be used for audit reporting and monitoring through an administrative console. The gateway operates as a virtualized instance and can run either on-premises or at a third-party hosted or managed service provider. The gateway can also function as a proxy, where it performs as a secure token service and point of policy enforcement, or in look-aside mode, where it passes on the identity logic to a third party to perform the transformations.

Cloud providers and security companies may together deliver a coordinated security approach that spans network, servers, databases, storage, and data, as well as connecting policies and controls across physical, virtual, and cloud infrastructure management activities.

3.7) **Choosing the right CSP**

Choosing a cloud service provider is complicated on many levels from the cloud delivery model and architecture to specific applications. Add to that the countless interdependencies and relationships, both technological and business-related, among vendors. To complicate matters, some companies offer not only software, but also hardware and services. Nevertheless, you must be vigilant about making sure the security you need to protect your data and platform are part of the offering.

At the highest level, you need to know if the cloud provider can provide evidence of data and platform protections for the services they provide. Once you are comfortable that your criteria can be met, you can establish measurable, enforceable SLAs to provide ongoing verification

SUMMARY:

Computing represents one of the most significant shifts in the information technology many of us are likely to see in our lifetimes. Reaching a point where computing functions as a point as a utility has great potential, promising innovations and we cannot yet imagine. Customers are both excited and nervous at the prospects of cloud computing and the difference it can have in the technology world. They are excited by the opportunities to reduce capital costs and a chance for themselves to divest them of the infrastructure management, and focus more in the core requirements of the companies. Most of all they are excited by the ability to align information technology with business strategies and other needs more readily. However customers are also very concerned about the risks of Cloud Computing if not properly secured, and the loss of direct control over systems for which they are nonetheless accountable.

Cloud fears largely stem from the perceived loss of control of sensitive data. Current control measures do not adequately address cloud computing's third-party data storage and processing needs. In our vision, we propose to extend control measures from the enterprise into the cloud. These measures should alleviate much of today's fear of cloud computing, and, we believe, have the potential to provide demonstrable business intelligence advantages to cloud participation.

REFERENCES :

[1] Amazon EC2 Crosses the Atlantic. http://aws. amazon.com/ about-aws/whats-new/2008/12 /10/ amazon-ec2-crosses-the-atlantic/.

[2] E.M. Stuart, J.D. John, Application and analysis of the virtual machine approachto information system security and isolation, in: Proceedings of the Workshopon Virtual Computer Systems, ACM, Cambridge, Massachusetts, United States,

[3] Amazon terms of use. http://aws.amazon.com/ agreement.

[4] An Information-Centric Approach to Information Security.http://virtualization.syscon.com/node/171 199.

[5] AOL apologizes for release of user search data. http://news.cnet.com/__ 2100-1030_3-6102793. html.

[6] Armbrust, M., Fox, A., Griffith, R. et al. Above the Clouds: A Berkeley View of Cloud Computing.UCB/EECS-2009-28, EECS Department, University ofCalifornia, Berkeley, 2009.

[7] Disaster-ProofingThe Cloud. http://www. forbes.com/ 2008 /11/24/cio-cloud-disaster-tech-cio-cx_dw_1125cloud.html.

[8] Security Evaluation of Grid Environments. https:// hpcrd.lbl.gov/ HEPCybersecurity/ HEP-Sec-Miller-Mar2005.ppt.

[9] Security issues with Google Docs. http://peekay.org/2009/03/26/security-issues-with-google-docs/.

[10] Frederick M. Avolio, Best Practices in Network Security, NetworkComputing,March 20, 2000, http://www. networkcomputing.com /1105/ 1105f2.html

[11] R. Sailer, X. Zhang, T. Jaeger, L. van Doorn, Design and implementation of a tcgbasedintegrity measurement architecture, in: Proceedings of the 13th USENIXSecurity Symposium, August 2004.

[12] Gabriel Mateescu, Wolfgang Gentzsch, Calvin J. Ribbens, Hybrid computingwhereHPC meets grid and cloud computing, Future Generation Computer Systems 27 (5) (2011) 440–453.

❑❑❑

Secure Key Pre-distribution in Wireless Sensor Networks Using Combinatorial Design and Traversal Design Based Key Distribution

Saba Khalid, Faiyaz Ahmad, Mohd. Rizwan Beg

Department of Computer Science and Engineering, Integral University, Lucknow,226026, India

Abstract - Security is an indispensable concern in Wireless Sensor Network (WSN) due to the presence of potential adversaries. For secure communication in infrastructureless sensor nodes various key predistribution have been proposed. In this paper we have evaluated various existing deterministic, probabilistic and hybrid type of key pre-distribution and dynamic key generation algorithms for distributing pair-wise, group-wise and network-wise keys and we have propose a key predistribution scheme using deterministic approach based on combinatorial design and traversal design which will improve the resiliency and achieve sufficient level of security in the network.This design can be used where large number of nodes are to be deployed in the WSN.

Keywords — *Sensor nodes(SN), Combinatorial design, Key pre-distribution scheme(KPS), Resiliency, Symmetric balanced incomplete block design(SBIBD), Traversal design*

I. INTRODUCTION

Sensor networks is a distributed adhoc network of collection of sensor nodes which are inexpensive devices having low battery power, low computation speed, limited memory capability and limited resources. Motivation of this paper is to evaluate the different key distribution solutions. On the basis of application types network architectures are classified such as distributed or hierarchical, communication styles such as pair-wise (unicast),group-wise (multicast) or network-wise (broadcast), security requirements such as authentication, confidentiality or integrity, and (iv) keying requirements such as pre-distributed or dynamically generated pair-wise, group-wise or network-wise keys. Key management services provide and manage the basic security material for satisfying the previously mentioned security services. In this paper we have presented a new KPS that uses combinatorial design and traversal design.

The rest of the paper is organised as in section II, deals with a brief background of combinatorial design theory. KPS is presented in section III. Section IV discusses and evaluates scalability issues and effects of node compromise in sensor networks. Finally in section V, the paper concludes with future work.

II. BACKGROUND: RELATED WORK

WSN consists of low power nodes which are randomly deployed and can effectively communicate to each other within a particular radio frequency range. According to their capability of communication nodes are classifieds as: (i) base stations (ii) cluster heads (iii) sensor nodes. For secure communication in SN keys can be either pre-distributed or online key exchange protocols can be used. Online key distribution scheme cannot be used as it requires public-key cryptography schemes which require more computational power. So the better option is to use key pre-distribution methods which are more secure and much faster.

Initially in WSN for security issues keys were distributed using a third trusted party called base stations (BS) proposed by Perrig et al. [1].Key distribution using this technique was not scalable and BS became a point of compromise. A KPS enables a SN to establish key without the use of BS. The simplest technique was to pre load the network with a single network wide key before deployment. But the disadvantage with this technique was that it was not scalable and comprise of a single node leads to compromise of all nodes in the network.

Inspired by the above idea Zhu. et. al [2] described pair wise key establishment scheme which relied on the assumption that no key will be compromised at the initial phase of sensor deployment and all sensors will erase their network wide key after initial phase. This scheme lacks scalability. The next step was using trivial pair-wise KPS but was limited in memory size and scalability. In the quest for security in KPS in SN Eschenauer and Gligor [3] proposed random key pre-

distribution scheme where tens to hundreds of keys were uploaded to SN before deployment. This scheme addresses unnecessary storage problem, initially a large key pool P is generated K keys are drawn randomly from P and stored in SN. This technique does not guarantee that any two nodes will be able to communicate directly. In order to establish a pairwise key two SN only needs to identity the common keys that they may share. If direct communication is not possible then a path needs to be established between two nodes. This makes communication power consuming and slower. Chan et al [4] proposed a modification of the scheme of [3] they extended this idea by allowing two sensors to setup a pair wise key only when they share at least q common keys.

This increased resiliency against node capture. Resiliency means the robustness under adverse conditions. Di Pietro et al. [5] applied a geometric random model for key pre-distribution, which further enhances the performance of previous KPSs. Hwang and Kim [6] proposed a method to improve performance of previous schemes by trading-off a very small number of isolated nodes.

In deterministic key pre-distribution, keys are placed in sensor nodes in a predetermined manner .The pioneering work of Camtepe *et al.* in [7] propose a deterministic pair wise key pre-distribution scheme based on expander graphs and projective planes. Lee and Stinson [8] used transversal designs, Chakrabarty, Maitra and Roy [9] used merging blocks constructed from transversal designs.

Here we have consider a deterministic key predistribution scheme based on combinatorial designs. The design finds application where a large number of sensor nodes are to be deployed. Also by suitably choosing the parameters of the design, it can be ensured that every pair of nodes within communication range can communicate directly, thus making communication efficient and less error-prone. The main advantage of this scheme is that it is resilient to selective node capture attack and node fabrication attack.

A: Theory on combinatorial design

Combinatorial design theory [7] is interested in arranging elements of a finite set into subsets to satisfy certain properties. A *Balanced Incomplete Block Design (BIBD)* is one of such designs. A *BIBD* is an arrangement of *v* distinct objects into *b* blocks such that each block contains exactly *k* distinct objects, each object occurs in exactly *r* different blocks, and every pair of distinct objects occurs together in exactly λ blocks. The design can be expressed as (v, k, λ), or equivalently (v, b, r, k, λ), where: $\lambda (v-1) = r (k-1)$ and b. k = v. r

A BIBD is called Symmetric BIBD or Symmetric Design when b = v. A Symmetric Design has four properties:

1. Every block contains k = r elements

2. Every element occurs in r = k blocks

3. Every pair of elements occurs in λ blocks

4. Every pair of blocks intersects in λ elements.

B: Projective plane

A Finite Projective Plane [9] consists of a finite set P of points and a set of subsets of P, called lines. For an integer n where $n \geq 2$, there are exactly n^2+ n + 1 point, and exactly n^2 + n + 1 line. If we consider lines as blocks and points as objects, then a Finite Projective Plane of order n is a Symmetric Design with parameters (n^2 + n + 1, n+ 1, 1) Finite Projective Plane of order n has four properties [8]:

1. Given any two distinct points, there is exactly one line incident with both of them.
2. Given any two distinct lines, there is exactly one point incident with both of them.
3. Every point has n+1 line through it.
4. Every line contains n+1 point.

A projective plane is therefore a symmetric (n^2 + n + 1, n+1, 1) block design.

A finite projective plane [8] exists when the order n is a power of a prime, i.e., for n = p1. It is conjectured that these are the only possible projective planes, but proving this remains one of the most important unsolved problems in combinatorics. The smallest finite projective plane is of order n = 2, consists of the configuration known as the Fano plane. This Fano plane, is denoted PG (2, 2).

A: Traversal design

A transversal design TD (k, n) [$k \geq 2$ and $n \geq 1$] is a triple (X, G, B) such that the following properties are satisfied:

1. X is a set of k. n elements called points,
2. G is a partition of X into k subsets of size n called groups,
3. B is a set of k-subsets of X called blocks,
4. Any group and any block contain exactly one common point, and
5. Every pair of points from distinct groups is contained in exactly one block.

III. COMBINATORIAL AND TRAVERSAL DESIGN BASED KPS

Combinatorial design provides an appropriate balance of key content in various sensor nodes. Using this strategy maximum number of nodes pair can communicate directly using pair wise common key. Transversal Design is such a combinatorial Design which offers a deterministic nature of key distribution. A pattern of key ids is seen in this type of distribution of keys. Lee and Stinson[8] first time proposed the application of Transversal Design for Key Pre-Distribution in WSN .The result is less communication with a balance distribution in the establishment of secure communication. As the property of TD yields maximum one pair wise key among node pair, therefore compromise of single key or node leads to the compromise of all the nodes and links having the same key and yields breaking of link and leads to victim nodes. Hence in adverse condition the resiliency is less due to the presence of single common key in the network.The term resiliency [16] refers to sustainability of the SN when some of its node have been compromised by the attacker. It is the security measure of a particular design and is measured by the parameter L(s): fraction of communication links compromise on compromise of randomly selected s number of node. Chakrabarti, Roy, and Maitra[9] has modified this scheme and proposed that instead of immediately considering each blocks as sensor node after distribution of keys using Transversal Design, a number of blocks can be merged to form a node yielding the probability of more than one common key between a pair of nodes. Therefore, during any adverse condition the probability of link breaking is least between a node pair. However, it increases memory space requirement which can be accommodated [9]. Additionally this scheme increases the resiliency. Selection of blocks for merging to form a node is purely random. Due to this randomness, the content of blocks in a node is random i.e. unpredictable.

During common key establishment between node pair an amount of communication cost O (x) is introduced, if number of blocks in each node is n. We have modified this part and proposed a deterministic scheme. In which we follow a peculiar rule for merging blocks to form a node. Since block selection is deterministic a pattern of blocks is formed in each node. Consequently, to uncover blocks for a specific node ,no extra communication cost is incurred during key establishment phase. Simulation and determination of the various parameters is performed. For simulation C Language is used as the platform.

A. ANALYSING THE APPROACH OF LEE AND STINSON'S SCHEME

Lee and Stinson have used the concept of TD for key predistribution in WSN as a result there is a pattern in key ids in each node.On studying and simulating the scheme provided by Lee and Stinson using C Language certain important parameters were studied like L(s) : Fraction of links which have been compromised due to the compromise of s number of nodes. The results obtain use (v, b, r, k) based transversal design, where v = 3232, b = 10201, r = 101, k = 32.

Maximum number of connection could be 104050200.

Number of initial links detected = 16070800.

Average number of common keys between node pair = 1.000000.

Therefore connectivity of the design is 0.164482, i.e. almost 16%.

The average value of L(s) = 0.3476, i.e. almost 34% where s =40.

TABLE I

showing outcome of L(s) for Lee and Stinson's scheme

S=4	L(s)= 0.0381
S=8	L(s)=0.0754
S=12	L(s)=0.0115
S=16	L(s)=0.1480
S=20	L(s)=0.1790
S=24	L(s)=0.2125
S=28	L(s)=0.2560
S=32	L(s)=0.2720
S=36	L(s)=0.3018
S=40	L(s)=0.3476

B. ANALYSING THE APPROACH OF CHAKRABARTI, ROY AND MAITRA'S SCHEME

According to Lee and Stinson's scheme, any node pair can share 0 or 1 key[8]. Merging of nodes to form a new node increases the number of common keys between a pair. Chakrabarti, Roy, and Maitra provide one scheme where they randomly choose x number of blocks and merged to form a new node. They have chosen the blocks in such a way that there will be no inter node connectivity. As they have chosen randomly,

for some cases they could not avoid the occurrence of inter node connectivity.

After forming a number of nodes they revised their scheme by introducing MOVE function to increase connectivity between different pairs in the network. MOVE increases the connectivity by exchanging blocks between maximum linked pair with zero linked pair.

On simulating this scheme the following parameters were studied L(S): Fraction of links get compromise on compromise of s number of nodes and Average number of common keys between a pair. The experiment result shows that the resiliency is much higher than the scheme provides by Lee and Stinson. But to store keys for each nodes need more storage. However, they have shown that consumed storage space is within the limits of a sensor node. The results obtained for various parameters are

- Maximum number of connection could be 3249974.
- Number of initial links detected is 3242103.
- Average number of common keys between a pair is 5.0195006.
- Therefore, connectivity of the design is 0.997589, i.e. almost 100%.
- The average value of L(s) = 0.0197, i.e. almost 2%, where s = 10 and equivalent to 40 blocks.

TABLE II

Outcome of L(s) for Chakrabarti, Ray and Matra's scheme

S=1	L(s)= 0.0010
S=2	L(s)= 0.0018
S=3	L(s)= 0.0028
S=4	L(s)= 0.0040
S=5	L(s)= 0.0062
S=6	L(s)= 0.0079
S=7	L(s)= 0.0100
S=8	L(s)= 0.129
S=9	L(s)= 0.0165
S=10	L(s)= 0.0197

C. *KEY DISTRIBUTION*

Chakrabarty, Roy and Maitra's scheme improves some parameters. However, it is observed that they have used randomly selected blocks to merge for forming node. Therefore, a particular node will be having no particular block id. On the time of shared key discovery between a pair of nodes, they have to broadcast all the block ids to the other nodes. This is yielding a communication cost O(x)[1], (x is the number of blocks to be merged to form a node) in addition to the request for communication which is O(1). Sending all the block ids cannot be avoided due to the randomness of the scheme. Observing this limitation, we propose a deterministic scheme for merging of block to form a node. The property of transversal design for arrangement of a set of elements into a number of subsets focuses the fact that the probability of repeating an element for consecutive blocks is much less. With such knowledge merging z ($1 \leq x \geq p$) number of blocks to form a node leads to much less probability for occurrence of intra-node repetition of same element. On the basis of this assumption, we considered x number of consecutive blocks for merging to form a node which helps to avoid any intra-node common key. This increases the connectivity of the entire network as well. Again as x number of consecutive blocks are merged, there is a pattern of block ids in a particular node. Therefore, to find out block ids for a particular node id there is no need to exchange block id which consumes an amount of communication effort. Nodes can themselves compute block ids of their counterparts. As this scheme is a deterministic, the communication cost is only O(1), that needs to request for communication by any of the node in the pair, which is much less than O(x). Note that the communication cost in this scheme is a constant value in comparison with scheme by Chakrabarti, Roy and Maitra where communication cost is a variable figure. On getting the node id of the requesting node, a node can easily determine the block ids of the other node which will take O(x) cost for computation time in average. After obtaining the block ids rest is to discover the shared keys, would take O(x^2log2^2r) time. Therefore, average computation cost for key establishment is O(x) + O(x^2log2^2r), i.e. O(x^2log2^2r), which is same as the scheme proposed by Chakrabarti, Roy, and Maitra. However, communication cost is much less which is one of the key requirements for these computational intensive devices. The algorithm for merging nodes is as follows:

/ Input: A block ids set*

Output: A node ids set

c=counter

t blocks= total number of blocks

u= number of blocks to be merged

*k= number of keys stored by each block */*

Start of blocks merging

C = 0;

For i = 0 to t blocks-1 do

For j = 0 to u-1 do

For s = 0 to k-1 do

Start

*Noderepository[i][j*k+s].1 = block[c][s].1;*

*/*Store first part of the key id*/*

*Noderepository[i][j*k+s].2= block[c][s].2;*

*/*Store second part of the key id*/*

End For

End For

C++;

End For

End of blocks merging

The experimented result are obtained using the design (v = 3232, b = 10201, r = 101, k = 32) and x = 4, is given below.

The total number of nodes which has formed is 2550 each having 128 number of keys.

Average number of common keys between two nodes 5.520075.

Maximum number of connection could be 3249975.

Number of initial links detected 2955867.

Therefore, connectivity of the design is 0.909465, i.e. almost 91%.

The average value of L(s) = 0.1552, i.e. almost 16%, where s = 10 and equivalent to 40 blocks.

TABLE III

Result of E(s) for proposed scheme

S=1	E(s)= 0.0080
S=2	E(s)= 0.0178
S=3	E(s)= 0.0300
S=4	E(s)= 0.0443
S=5	E(s)= 0.0578
S=6	E(s)= 0.0800
S=7	E(s)= 0.0960
S=8	E(s)= 0.1125
S=9	E(s)= 0.1455
S=10	E(s)= 0.1557

D. KEY EXCHANGE

Any pair wishes to communicate with each other send a request message to its counterpart, which then including the sender discovers the common key between them. According to the proposed scheme, they don't need to send any extra information. They generate the block ids of the others using the above algorithm which needs the node id only of the other node. On discovering the block ids, they can compare all the blocks with their own blocks for finding any common key using the algorithm proposed by Lee and Stinson. After discovering the common key, if any, they can start communication using that key. In case of a pair which does not have any common key, they have to generate a key temporarily and need to exchange through one or more intermediate nodes. This process is referred as path key establishment.

IV. COMPUTATIONAL RESULTS

When we compare our scheme wee see that our scheme requires computation of O(1) to calculate shared keys. This is because our scheme broadcast only node identifier whereas other schemes have to share key identifiers. Though scheme proposed by Chakrabarti, Roy,and Maitra consumes a variable communication cost O(x), where x is the number of blocks to be merge to form a node. Again, though the scheme proposed by Lee and Stinson consumes O(1) as the communication cost, it still suffers from less Resiliency. Computation for key discovery is same i.e. O(x^2log2^2r) in this scheme as well as for Chakrabarti, Roy and Maitra and Lee & Stinson. The average number of common keys in each pair of node is almost 5 in this scheme as well as in [9], whereas scheme proposed by Lee and Stinson[8] has only 1 key. This is the main advantage of merging blocks to form node. Connectivity of this scheme is almost 91% which is almost same with the scheme proposed by [9]. Nevertheless, connectivity of the proposed scheme is much better than the scheme proposed by [8]. The resiliency is best in Chakrabarti, Roy and Maitra's scheme. Given the limited memory space and battery constraint our scheme shows reasonable resilience and better node connectivity especially when a large number of nodes have been compromised.

Node pair within a radio frequency range can communicate with each other, provided they have a common key between them. In probabilistic schemes this is not possible as nodes are chosen randomly. We see in our deterministic scheme any two nodes share at least one key. So there is full connectivity in the network.

V. SECURITY ISSUES IN WSN

WSN inherits security problems due to :-

(i) Wireless nature of communication,

(ii) Limitation of capability of individual sensor nodes,

(iii) Large size of the sensor networks,

(iv) Unknown and dynamic network topology, and

(v) Easy chance of physical attack.

This results a challenge to design any efficient key management scheme. We have tried to evaluate scheme in terms of its resilience against node collusion and selective node capture attack. If a node is directly involved in node collusion, e.g., because of being captured by an outside adversary or re-programmed to do harm to the whole network, we say the node is *compromised*. We have tried to answer the question that when certain nodes are compromised, how much could they influence the rest of the network if their key information has been retrieved and analyzed.

Our scheme is resilient to selective node capture attack , during this attack the attacker comprise those nodes whose keys have not already been compromised. Amid shared key discovery phase only node identifiers are broadcasted, key identifiers are not exchanged. Hence attacker at any stage cannot know which key identifiers are present in which node. Thus attacker cannot gain any information using this attack.

VI. CONCLUSION

On studying and comparing different schemes we find that merging of blocks to form node improves a number of parameters such as resiliency, average number of common key between a node pair, connectivity of the network etc. Only limitation is that the memory usage is significantly large, however this requirement is easily adaptable. Many existing schemes including [9] merge randomly therefore communication cost for key discovery is more and equivalent to O(x), which we have tried to reduce by proposing a deterministic scheme. This is one of the major requirements for a wireless sensor network.

Future direction of work will be to further study our scheme from other perspectives, such as computational overheads and investigate approaches to increase resilience by revising the merging strategy.

ACKNOWLEDGMENT

The authors would like to thank Mr. Shish Ahmad for his valuable comments and persistent efforts.

REFERENCE :

[1] Perrig, R. Szewczyk, V. Wen, D. Cullar, and J. D. Tygar. SPINS: Security protocols for sensor networks. In Proc. of MOBICOM, 2001

[2] S. Zhu, S. Setia and S. Jajodia. LEAP: Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks. In. Proc. of the ACM CCS Conference, pp. 62-72. 2003

[3] L. Eschenauer and V. D. Gligor. A key-management scheme for distributed sensor networks. In Proc. of the 9th ACM CCS conference, pp. 41 – 47, 2002

[4] H. Chan, A. Perrig, and D. Song. Random key predistribution schemes for sensor networks. In Proc. of the IEEE Symposium on Security and Privacy, p. 197, 2003M.

[5] R. Di Pietro, L. V. Mancini, A. Mei, A. Panconesi. Connectivity Properties of Secure Wireless Sensor Networks. In Proc. of the 2nd ACM SASN workshop, pp. 53 – 58. 2004.

[6] J. Hwang and Y. Kim. Revisiting random key predistribution schemes for wireless sensor networks. In Proc. Of the 2nd ACM SASN workshop, pp. 43 – 52. 2004

[7] .S. A. Camtepe and Bülent Yener. Combinatorial Design of Key Distribution Mechanisms for Wireless Sensor Networks. In Proc. of Computer Security- ESORICS, Springer-Verlag, LNCS 3193, 2004, pp 293-308.

[8] J. Lee, and D. R. Stinson. Deterministic Key Predistribution Schemes for Distributed Sensor Networks. In Proc. 11th International Workshop, SAC 2004, pp. 294-307.

[9] D. Chakrabarti, S. Maitra, and B. K. Roy. A key predistribution scheme for wireless sensor networks: merging blocks in combinatorial design. Int. J. Inf. Sec., 5(2):105–114, 2006.

[10] David Sánchez Sánchez, Heribert Baldus. A Deterministic Pairwise Key Pre-distribution Scheme for Mobile Sensor Networks .In Proc of the First International Conference on Security and Privacy for Emerging Areas in Communications Networks,2005, IEEE.

[11] H. Shafiei, A. Mehdizadeh, A. Khonsari and M. Ould-Khaoua. A Combinatorial Approach for Key-Distribution in Wireless Sensor Networks. In Proc of the IEEE "GLOBECOM" 2008.

[12] SushmitaRuj, Jennifer Seberry and Bimal Roy. Key Predistribution Schemes Using Block Designs in Wireless Sensor Networks. In proc of International Conference on Computational Science and Engineering, 2009.

[13] W. Stallings, Cryptography and Network Security- PrinCiples and Practices, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2003.

[14] Yingshu Li, My T. Thai, Weili Wu. Wireless Sensor Networks and Applications.Springer,2008

[15] Anupam Pattanayak, B. Majhi. Key Predistribution Schemes in Distributed Wireless Sensor Network using Combinatorial Designs Revisited. Cryptology eprint Archive. Report 2009/131. 2009

[16] Anupam Pattnayak, "Deterministic Merging of Blocks in Combinatorial Design based Key Predistribution in Distributed Wireless Sensor Netwo," M. Eng. thesis, National Institute of Technology, Orissa, India, May, 2009.

[17] Subhasish Dhal, "Application of Traversal Design and Secure Path Key Establishment for Key Pre-Distribution in WSN", M.Eng. thesis, National Institute of Technology, Orissa, India, May, 2009 .

Impulse Noise Removal From Color Image Sequences Using Fuzzy Filter

M.V.Phani Kumar & T.Venkata Lakshmi
Dept. of ECE, GEC, Gudlavalleru, Dist. Krishna, Andhra Pradesh, India

Abstract - Digital image processing is a subset of the electronic domain wherein the image is converted to an array of small integers, called pixels, representing a physical quantity such as scene radiance, stored in a digital memory, and processed by computer or other digital hardware. In this paper, a new fuzzy filter for the removal of random impulse noise in color video is presented. By working with different successive filtering steps, a very good tradeoff between detail preservation and noise removal is obtained. In order to preserve the details as much as possible, the noise is removed step by step. Pixels that are detected as noisy are filtered, the others remain unchanged. The detection of noisy color components is based on fuzzy rules in which information from spatial and temporal neighbors as well as from the other color bands is used. The experiments show that the proposed method outperforms other state-of-the-art filters both visually and in terms of objective quality measures such as the mean absolute error (MAE), the peak-signal-to-noise ratio (PSNR) and the normalized color difference (NCD).

Keywords-Digital Image Processing (DIP); fuzzy logic (FL); Peak-signal-to-noise-ratio (PSNR); normalized color difference (NCD); mean absolute error (MAE).

I. INTRODUCTION

Generally, noise in signal processing is interpreted as 'unwanted signals'. However, in the context of image processing, noise is termed as the displacement of the signal intensities from their original values. A fundamental problem of image analysis is to effectively remove noise from an image while keeping its fundamental structure constituting of edges, corners, etc., intact. The nature of the noise removal problem depends on the type of the noise corrupting the image. However, the images are likely to be corrupted by noise due to bad acquisition, transmission or recording. Such degradation negatively influences the performance of many image processing techniques and a preprocessing module to filter the images is often required.

The impulse noise (or salt and pepper noise) is caused by sharp, sudden disturbances in the image signal; its appearance is randomly scattered white or black (or both) pixels over the image. Noise filtering can be viewed as removing the noise from the corrupted image and smoothen it so that the original image can be viewed. Noise filtering can be viewed as replacing every pixel in the image with a new value depending on the fuzzy based rules. Ideally, the filtering algorithm should vary from pixel to pixel based on the local context.

The objective of the paper is to give a new better, faster and efficient solution for removing the noise from the corrupted images. The main point under consideration is that the noise-free pixels must remain unchanged. The main focus will be on:

1. Removal of the noise from the test image.
2. Noise free pixels must remain unchanged.
3. Edges must be preserved.
4. Improve the contrast

Fuzzy set theory was introduced by Zadeh in 1965 and is a generalization of classical set theory. A classical crisp set over a universe X can be modeled by a X-{0,1}mapping (characteristic function): an element belongs to x ϵ X belongs the set or does not belong to it. Fuzzy sets are now modelled as X-[0, 1] mappings (membership functions). So the characteristic function is extended to a membership function where also membership degrees between zero and one are allowed. An element x ϵ X which allows a more gradual transition between belonging to and not belonging to. Most filters in literature, which are developed for video, are intended for sequences corrupted by additive Gaussian noise. Only few video filters for the impulse noise case can be found. However, several impulse noise filters for still images exist. The best known among them are the median based rank-order filters. But also some fuzzy techniques can be found. Such 2-D filters could be used to filter each of the frames of a video successively. However, temporal inconsistencies will arise due to the neglection of the temporal correlation between successive frames. A better alternative would be to use 3-D filtering windows, in which also pixels from neighboring frames are taken into account. The main problem in using neighboring frames is motion between them. Using pixels at

corresponding spatial positions in neighboring frames for noise removal may introduce ghosting artifacts in the presence of camera and object motion. In the method proposed in this paper, we will therefore only in non-moving areas assign a temporal impulse between two corresponding spatial positions to noise (detection phase) and for the replacement of a noisy pixel (filtering phase) motion compensation will be applied to find the most reliable pixel in the previous frame. Filters for grayscale images could be used for color images by applying them on each of the color bands of the image separately.

II. THE PROPOSED ALGORITHM

The proposed filtering framework consists of three successive filtering steps as depicted in Fig. 1. By removing the noise step by step, the details can be preserved as much as possible. Indeed, if a considerable part of the noise has already been removed in a previous step, and more noise-free neighbours to compare to be available, it will be easier to distinguish noise from small details.

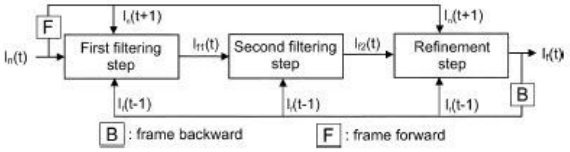


Fig. 1. Overview of the different steps in the proposed algorithm.

In the first step (with output denoted by I_{f1}), we calculate for each pixel component a degree to which it is considered noise-free and a degree to which it is considered noisy. If the noisy degree is larger than the noise-free degree, the pixel component is filtered, otherwise it remains unchanged. The determination of both degrees is mainly based on temporal information (comparison to the corresponding pixel component in the previous frame). Note, however, that only in non-moving areas can large temporal differences be assigned to noise. In areas where there is motion, such differences might also be caused by that motion. As a consequence, and as can be seen in Fig. 2, impulses in moving areas will not always be detected in this step. They can, however, be detected in the second step (output I_{f2}). Analogously as to the first step, again a noise-free degree and a noisy degree are calculated. However, the detection is now mainly based on color information. A pixel component can be seen as noisy if there is no similarity to its (spatio-temporal) neighbors in the given color, while there is in the other color bands. The third step (output I_f), finally, removes the remaining noise and refines the result by using as well temporal as spatial and color information. For example, homogeneous areas can be refined by removing small impulses that are relatively large in that region, but are not large enough to be detected in detailed regions and that thus have not been detected yet by the previous general detection steps. The results of the different successive filtering steps are illustrated for the 20th frame of the "Salesman" sequence in Fig. 2.

A) First Filtering Step

1) Detection: In this detection step, we calculate for each of the components of each pixel a degree to which it is considered noise-free and a degree to which it is thought to be noisy. A component for which the noisy degree is larger than the noise free degree, i.e., that is more likely to be noisy than noise-free, will be filtered. Other pixel components will remain unchanged.

Fig. 2 : The original 20th frame of the "Salesman" sequence (a), the frame corrupted by 20% random impulse noise (b) (PSNR=15.05 dB) and the result after the first (c) (PSNR=23.72 dB), second (d) (PSNR=29.42 dB) and refinement step (e) (PSNR=36.78 dB) respectively.

The noise-free degree and the noisy degree are determined by fuzzy rules as follows. We consider a pixel component to be noise-free if it is similar to the corresponding component of the pixel at the same spatial location in the previous or next frame and to the corresponding component of two neighboring pixels in the same frame. In the case of motion, the pixels in the

previous frames can not be used to determine whether a pixel component in the current frame is noise-free. Therefore, more confirmation (more similar neighbors or also similar in the other color components) is wanted instead. For the noise-free degree of the red component (and analogously for the other components), this is achieved by the following fuzzy rule. To represent the linguistic value *large positive* in the above rule, a fuzzy set is used, with a membership function as depicted in Fig. 3. Those operators are simple in use and yielded the best results, but the difference compared to the results for another choice of operators is neglectible.

Fuzzy Rule 1: IF($|I_n^R(x,y,t) - I_f^R(x,y,t-1)|$ is NOT LARGE POSITIVE OR $I_n^R(x,y,t) - I_n^R(x,y,t+1)$ is NOT LARGE POSITIVE) AND there are two neighbors $(x+k, y+l, t)$ $(-2 \leq k,l \leq 2$ and $(k,l) \neq (0,0))$ for which $|I_n^R(x,y,t) - I_n^R(x+k,y+l,t)|$ is NOT LARGE POSITIVE)

OR (there are four neighbors $(x+k, y+l, t)$ $(-2 \leq k,l \leq 2$ and $(k,l) \neq (0,0))$ for which $|I_n^R(x,y,t) - I_n^R(x+k,y+l,t)|$ is NOT LARGE POSITIVE OR (there are two neighbors

$(x+k, y+l, t)$ $(-2 \leq k,l \leq 2$ and $(k,l) \neq (0,0))$ for which $I_n^R(x,y,t) - I_n^R(x+k,y+l,t)$ is NOT LARGE POSITIVE AND $(|I_n^G(x,y,t) - I_n^G(x+k,y+l,t)|$ OR $I_n^B(x,y,t) - I_n^B(x+k,y+l,t)$ are NOT LARGE POSITIVE)))

THEN the red component $I_n^R(x,y,t)$ is considered NOISE-FREE.

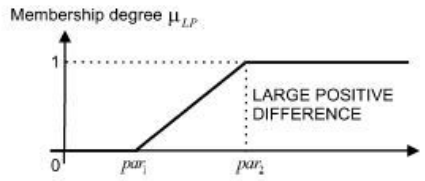


Fig. 3. The membership function μ_{LP} of the fuzzy set *large positive*.

The outcome of the rule, i.e., the degree to which the red component of the pixel at position (x, y, t) is considered noise-free, is determined as the degree to which the antecedent in the fuzzy rule is true. For the conjunctions (AND), disjunctions (OR) and negations (NOT) in fuzzy logic, triangular norms, triangular conorms and involutive negators are used. Analogously, a degree to which the component of a pixel is considered noisy is calculated. In this step, we consider a pixel component to be noisy if the absolute difference in that component is large positive compared to the pixel at the same spatial location in the previous frame and if not for five of its neighbours the absolute difference in this component and one of the other two color bands is large positive compared to the pixel at the same spatial location in the previous frame (which means that the difference is not caused by motion). Further, we also want a confirmation either by the fact that in this color band, there is a direction in which the differences between the considered pixel and the two respective neighbours in this direction are both large positive or large negative and if the absolute difference between those two neighbours is not large positive (i.e., there is an impulse between two pixels that are expected to belong to the same object) or by the fact that there is no large difference between the considered pixel and the pixel at the same spatial location in the previous frame in one of the other two color bands. For the red component (and analogously the other components) this leads to the following fuzzy rule.

Fuzzy Rule 2: IF($I_n^R(x,y,t) - I_f^R(x,y,t-1)$ is LARGE POSITIVE AND NOT (for five neighbors $(x+k, y+l, t)$ $(-2 \leq k,l \leq 2$ and $(k,l) \neq (0,0))$ $I_n^R(x+k,y+l,t) - I_f^R(x+k,y+l,t-1)$ is LARGE POSITIVE AND $(|I_n^G(x+k,y+l,t) - I_f^G(x+k,y+l,t-1)$ OR $I_n^B(x+k,y+l,t) - I_f^B(x+k,y+l,t-1)$ is LARGE POSITIVE)))

AND ((in one of the four directions (the differences $I_n^R(x,y,t) - I_n^R(x+k,y+l,t)$ AND $I_n^R(x,y,t) - I_n^R(x-k,y-l,t)$ $((k,l) \in \{(-1,-1),(-1,0),(-1,1),(0,1)\})$ are both LARGE POSITIVE OR both LARGE NEGATIVE) AND the absolute difference $I_n^R(x+k,y+l,t) - I_n^R(x-k,y-l,t)|$ is NOT LARGE POSITIVE) OR $(|I_n^G(x,y,t) - I_f^G(x,y,t-1)|$ is NOT LARGE POSITIVE OR $|I_n^B(x,y,t) - I_f^B(x,y,t-1)|$ is NOT LARGE POSITIVE))

THEN the red component $I_n^R(x,y,t)$ is considered noisy.

Analogously to the linguistic term *large positive*, also *large negative* is represented by a fuzzy set, characterized by the membership function given in Fig. 4.

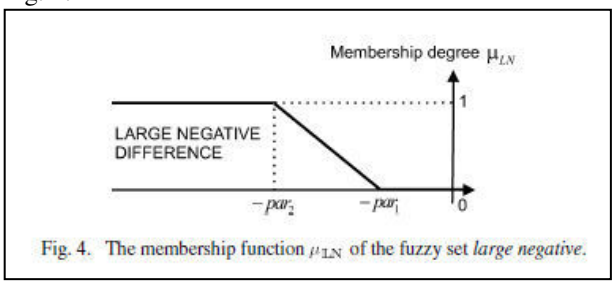


Fig. 4. The membership function μ_{LN} of the fuzzy set *large negative*.

2) Filtering: In this subsection, we discuss the filtering for the red color band. The filtering of the other color bands is analogous. We decide to filter all red pixel components that are considered more likely to be noisy than noise-free. The red components of the other pixels remain unchanged to avoid the filtering of noise-free pixels (that might have been uncorrectly assigned a low noisy degree, but for which the high noise-free degree assures us that it is noise-free) and thus detail loss. On the other hand, noisy pixel components might remain unfiltered due to an uncorrect high noise-free degree,

but those pixels can still be detected in the next filtering step.

B) Second Filtering Step

In our aim to preserve the details as much as possible, the noise is removed in successive steps. In this step, the noise is detected based on the output of the previous step. Also in this second filtering step, a degree to which a pixel component is expected to be noise-free and a degree to which a pixel component is expected to be noisy, is calculated. In the calculation of those degrees, we now take into account information from the other color bands.

A color component of a pixel is considered noise-free if the difference between that pixel and the corresponding pixel in the previous frame is not large in the given component and also not large in one of the other two color components. It is also considered noise-free if there are two neighbours for which the difference in the given component and one of the other two components are not large. So, the other color bands are used here as a confirmation for the observations in the considered color band to make those more reliable.

For the red component (and analogously the other color components), this gives the following fuzzy rule.

Fuzzy Rule 3: IF($I^R_{f_1}(x,y,t) - I^R_f(x,y,t-1)$ is NOT LARGE POSITIVE AND($I^G_{f_1}(x,y,t) - I^G_f(x,y,t-1)$ is NOT LARGE POSITIVE OR $|I^B_{f_1}(x,y,t) - I^B_f(x,y,t-1)$ is NOT LARGE POSITIVE)

OR(for two neighbors $(x+k,y+l,t)$ $(-1 \le k,l \le 1$ and $(k,l) \ne (0,0))$ $|I^R_{f_1}(x,y,t) - I^R_f(x+k,y+l,t)$ is NOT LARGE POSITIVE AND($|I^G_{f_1}(x,y,t) - I^G_f(x+k,y+l,t)$ is NOT LARGE POSITIVE OR $|I^B_{f_1}(x,y,t) - I^B_f(x+k,y+l,t)$ is NOT LARGE POSITIVE))

THEN the red component $I^R_{f_1}(x,y,t)$ is considered NOISE-FREE.

The degree to which the red component of the pixel at position (x,y,t) is considered noise-free, is then given by

$$\mu^R_{2,\text{noisefree}}(x,y,t) = \max(\zeta(x,y,t), \eta(x,y,t))$$

where

$$\begin{aligned}\zeta(x,y,t) &= \min(1-\mu_{\text{LP}}(|I^R_{f_1}(x,y,t) - I^R_f(x,y,t-1)|),\\ &\quad \max(1-\mu_{\text{LP}}(|I^G_{f_1}(x,y,t) - I^G_f(x,y,t-1)|),\\ &\quad 1-\mu_{\text{LP}}(|I^B_{f_1}(x,y,t) - I^B_f(x,y,t-1)|)))\end{aligned}$$

and $\eta(x,y,t)$ is the second largest element in the set

$$\begin{aligned}&\{\min(1-\mu_{\text{LP}}(|I^R_{f_1}(x,y,t) - I^R_f(x+k,y+l,t)|),\\ &\quad \max(1-\mu_{\text{LP}}(|I^G_{f_1}(x,y,t) - I^G_f(x+k,y+l,t)|),\\ &\quad 1-\mu_{\text{LP}}(|I^B_{f_1}(x,y,t) - I^B_f(x+k,y+l,t)|)))\\ &\quad \| -1 \le k,l \le 1 \text{ and } (k,l) \ne (0,0)\}.\end{aligned}$$

A pixel component is considered noisy if there are three neighbours that differ largely in that component, but are similar (not a large difference) in the other two components. It is also considered noisy if in the considered color band, its value is larger or smaller than the component values of all its neighbours, and this is not the case in both of the other color bands.

For the red component of a pixel (and analogously for the other components), this corresponds to the following fuzzy rule.

Fuzzy Rule 4: IF (for three neighbors $(x+k,y+l,t)$ $(-1 \le k,l \le 1$ and $(k,l) \ne (0,0))$ $|I^R_{f_1}(x,y,t) - I^R_f(x+k,y+l,t)|$ is LARGE POSITIVE AND $|I^G_{f_1}(x,y,t) - I^G_f(x+k,y+l,t)$ is NOT LARGE POSITIVE AND $|I^B_{f_1}(x,y,t) - I^B_f(x+k,y+l,t)|$ is NOT LARGE POSITIVE)

OR (((for all neighbors $(x+k,y+l,t)$ $(-1 \le k,l \le 1$ and $(k,l) \ne (0,0))$ $|I^R_{f_1}(x,y,t) - I^R_f(x+k,y+l,t)|$ is LARGE POSITIVE) OR (for all neighbors $(x+k,y+l,t)$ $(-1 \le k,l \le 1$ and $(k,l) \ne (0,0))$ $|I^R_{f_1}(x,y,t) - I^R_f(x+k,y+l,t)|$ is LARGE NEGATIVE)) AND NOT (((for all neighbors $(x+k,y+l,t)$ $(-1 \le k,l \le 1$ and $(k,l) \ne (0,0))$ $|I^G_{f_1}(x,y,t) - I^G_f(x+k,y+l,t)|$ is LARGE POSITIVE) OR (for all neighbors $(x+k,y+l,t)$ $(-1 \le k,l \le 1$ and $(k,l) \ne (0,0))$ $|I^G_{f_1}(x,y,t) - I^G_f(x+k,y+l,t)|$ is LARGE NEGATIVE)) AND ((for all neighbors $(x+k,y+l,t)$ $(-1 \le k,l \le 1$ and $(k,l) \ne (0,0))$ $|I^B_{f_1}(x,y,t) - I^B_f(x+k,y+l,t)|$ is LARGE POSITIVE) OR (for all neighbors $(x+k,y+l,t)$ $(-1 \le k,l \le 1$ and $(k,l) \ne (0,0))$ $|I^B_{f_1}(x,y,t) - I^B_f(x+k,y+l,t)|$ is LARGE NEGATIVE))))

C) Third Filtering Step

The result from the previous steps is further refined based on temporal, spatial and color information. Namely, the red component (and analogously the green and blue component) of a pixel is refined. Very small impulses might not have been detected by the algorithm. In homogeneous areas however, such impulses might be relatively large and can be detected more easily.

III. EXPERIMENTAL RESULTS

To be able to judge the performance of the proposed method, we will use the mean absolute error (MAE), the peak-signal-to noise-ratio (PSNR) and the normalized color difference (NCD) as objective measures of similarity and dissimilarity between a filtered frame and the original one, each containing m rows and n columns of pixels.

The MAE is given by

$$\mathrm{MAE}(I_o(t), I_f(t)) = \frac{\sum_{c\in\{R,G,B\}} \sum_{x=1}^{m} \sum_{y=1}^{n} |I_o^c(x,y,t) - I_f^c(x,y,t)}{3 \cdot n \cdot m}.$$

The lower the MAE, the more similar (less dissimilar) the images.

The MSE value is defined as

$$\mathrm{MSE}(I_o(t), I_f(t)) = \frac{\sum_{c\in\{R,G,B\}} \sum_{x=1}^{m} \sum_{y=1}^{n} (I_o^c(x,y,t) - I_f^c(x,y,t))^2}{3 \cdot n \cdot m},$$

The PSNR value is defined as

$$\mathrm{PSNR}(I_o(t), I_f(t)) = 10 \cdot \log_{10} \frac{S^2}{\mathrm{MSE}(I_o(t), I_f(t))},$$

Where S denotes the maximum possible value of a pixel component (here S= 255). The higher the PSNR value, the more similar (less dissimilar) the images.

Finally, the NCD, between an original and a filtered frame, is calculated as

$$\mathrm{NCD}(I_o(t), I_f(t)) = \frac{\sum_{x=1}^{m} \sum_{y=1}^{n} \left\| I_o^{\mathrm{LAB}}(x,y,t) - I_f^{\mathrm{LAB}}(x,y,t) \right\|}{\sum_{x=1}^{m} \sum_{y=1}^{n} \left\| I_o^{\mathrm{LAB}}(x,y,t) \right\|}$$

where $\| \cdot \|$ is the Euclidean norm and $I_o^{\mathrm{LAB}}(x,y,t)$ and $I_f^{\mathrm{LAB}}(x,y,t)$ respectively denote the $L^*a^*b^*$-transform [42] of the original and the filtered frame. The lower the NCD value, the more similar (less dissimilar) the images.

A) Parameter Selection

First the parameters and that determine the membership functions μ_{LP} and μ_{LN} in Figs. 3 and 4 are determined. To do this, we have fixed the window sizes $W_1$and $W_2$of the pixel neighborhood and the search region in the filtering as W_1=2(5x 5 neighborhood) and W_2=5 (11x11 search region) and we have let the parameters par_1 and par_2 run over range of possible values. The obtained values, which we will also use in the remaining experiments, are (par_1, par_2) = (20, 31) as can be seen in (Table I).

Next, the windows sizes $W_1$and W_2 are set. For the above selected parameter values for and, we now let the parameters $W_1$and W_2 run over a range of possible values. As can be seen in Table II, from the couple (W_1,W_2) = (2, 7) on, the arithmetic mean of the PSNR values of the nine test sequences hardly increases. Although we have focused in this paper on the noise filtering capability of the filter and not on its complexity, it should be mentioned that most of the computation time needed by the method goes to the filtering of detected pixels, i.e., the search for the best matching block.

TABLE I
DETERMINATION OF THE PARAMETERS par_1 AND par_2 (ARITHMETIC MEAN OF THE AVERAGE PSNR (dB) VALUES AROUND THE MAXIMUM)

$par_1 \backslash par_2$	29	30	31	32	33
18	32.38	32.39	32.40	32.38	32.37
19	32.39	32.41	32.41	32.40	32.39
20	32.39	32.41	32.42	32.40	32.39
21	32.39	32.40	32.41	32.40	32.39
22	32.39	32.40	32.41	32.40	32.39

Although we have focused in this paper on the noise filtering capability of the filter and not on its complexity, it should be mentioned that most of the computation time needed by the method goes to the filtering of detected pixels, i.e., the search for the best matching block. The size of a block (the number of pixels that has to be handled for each block) and the size of the search region (the number of blocks to which a given block should be compared) increases quadratic with respect to respectively W_1 and W_2.

TABLE II
DETERMINATION OF THE PARAMETERS W_1 AND W_2 (ARITHMETIC MEAN OF THE AVERAGE PSNR (dB) VALUES)

$W_1 \backslash W_2$	5	6	7	8	9	10	11
1	31.42	31.42	31.41	31.38	31.33	31.31	31.27
2	32.42	32.49	32.57	32.57	32.55	32.55	32.54
3	32.45	32.54	32.63	32.64	32.64	32.64	32.64
4	32.46	32.56	32.65	32.67	32.67	32.67	32.67
5	32.35	32.45	32.55	32.57	32.57	32.58	32.58

B) Comparison to Other State-of-the-Art Filters

In this subsection, the performance of the proposed method is compared to that of the adaptive vector median filter (AVMF) from , the video adaptive vector directional median filter (VAVDMF) with 3-D filtering window from and the 2-D fuzzy impulse noise reduction

method for color images (INRC) .The adaptive vector median filter orders the pixels (color vectors) in the 3-D filtering window based on increasing accumulated (Euclidean) distance to the other pixels in the window. If the Euclidean distance between the central pixel in the window and the mean of a given number of vectors that have the lowest accumulated distance, is greater than a given threshold, then the central pixel is filtered as the pixel with the lowest accumulated distance, otherwise, it remains unchanged. In the video adaptive vector directional median filter the vectors are first ordered based on increasing angular distance. If the absolute distance between the central pixel in the window and the mean of a given number K of vectors that have the lowest accumulated angular distance, is greater than a given threshold, then the central pixel is filtered as the pixel with the lowest accumulated absolute distance (magnitude), otherwise, it remains unchanged.

To show that the proposed filter takes real advantage from the temporal information, we have also compared the proposed filter to the 2-D fuzzy impulse noise reduction method for color images. The INRC filter outperforms all other compared state-of-the-art 2-D methods and can thus be accepted as a good representative for the 2-D impulse noise filters. Further, this filter is also a representative of a non-vector-based filter, in which the color bands are filtered separately. However, in the detection of noisy pixel components, also information from the other components is used.

C) Some Notes on the Complexity

As shown in the previous subsection, the proposed filtering frame work outperforms other state-of-the-art filters for video corrupted by random impulse noise. However, it also needs to be said that in the development of our filtering framework, the main focus was the filtering result and not the complexity, as it was more the case for the compared methods. Note that the largest computational cost of the proposed filter can be attributed to the block matching in the filtering stage. Since only pixels that are detected as noisy are filtered, the number of pixels that are filtered, and thus the running time of the algorithm, will increase with increasing noise level. As an illustration, Table III gives the running time for the processing of the "Salesman" sequence corrupted by different noise levels by the proposed algorithm.

The algorithm was implemented in Matlab in combination with the mex-function and executed on an Intel® Xeon® CPU X3360 @ 2.83 GHz. Some suggestions to reduce the computation time needed by the proposed filter could be the following.

TABLE III
AVERAGE RUNNING TIME (SECONDS PER FRAME) FOR THE PROCESSING OF THE "SALESMAN" SEQUENCE

	% random impulse noise						
	0	5	10	15	20	25	30
INRC	0.65	0.67	0.69	0.71	0.73	0.75	0.76
VAVDMF	7.77	7.72	7.67	7.65	7.65	7.59	7.57
AVMF	1.21	1.21	1.22	1.21	1.21	1.22	1.22
Proposed	2.54	4.76	6.92	8.98	10.95	12.90	14.84

First, the block matching in the filtering stage could be sped up by using fast motion estimation techniques. Next, for higher noise levels, it might be useful to do the block matching for blocks of pixels and to filter each of the noisy pixels in the blocks at the same time instead of applying the block matching for each noisy pixel separately. Further, in each of the successive steps in the algorithm, the detection and filtering of a pixel does not depend on the detection and filtering of the other pixels in the frame, such that the algorithm could be further sped up by performing this detection and filtering for several pixels in parallel.

IV. CONCLUSION

In this paper, we have presented a new filtering framework for color videos corrupted with random valued impulse noise. In order to preserve the details as much as possible, the noise is removed step by step. The detection of noisy color components is based on fuzzy rules in which information from spatial and temporal neighbours as well as from the other color bands is used. Detected noisy components are filtered based on block matching where a noise adaptive mean absolute difference is used and where the search region contains pixels blocks from both the previous and current frame. The experiments showed that the proposed method outperforms other state-of-the-art methods both in terms of objective measures such as MAE, PSNR and NCD and visually.

REFERENCE

[1] E. Abreu, M. Lightstone, S. K. Mitra, and K. Arakawa, "A new efficient approach for the removal of impulse noise from highly corrupted images," *IEEE Trans. Image Process.*, vol. 5, no. 6, pp. 1012–1025,1996.

[2] R. H. Chan, C. Hu, and M. Nikolova, "An iterative procedure for removing random-valued impulse noise," *IEEE Signal Process. Lett.*, vol.. 11, no. 12, pp. 921–924, 2004.

[3] [3] S. M. M. Rahman, M. O. Ahmad, and M. N. S. Swamy, "Video denoising based on inter-

frame statistical modelling of wavelet coefficients," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 2, pp.187–198, 2007.

[4] L. Jovanov, A. Pizurica, V. Zlokolica, S. Schulte, P. Schelkens, A.Munteanu, E. E. Kerre, and W. Philips, "Combined wavelet-domain and motion-compensated video denoising based on video codec motion estimation methods," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 417–421, 2009.

[5] H. B. Yin, X. Z. Fang, Z. Wei, and X. K. Yang, "An improved motion-compensated 3-D LLMMSE filter with spatio-temporal adaptive filtering support," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 12, pp. 1714–1727, 2007.

[6] L. Guo, O. C. Au, M. Ma, and Z. Liang, "Temporal video denoising based on multihypothesis motion compensation," *IEEE Trans. Circuits Syst. Video Technol*, vol. 17, no. 10, pp. 1423–1429, 2007.

[7] T. Mélange, M. Nachtegael, E. E.Kerre, V. Zlokolica, S. Schulte, V. De Witte, A. Pizurica, and W. Philips, "Video denoising by fuzzy motion and detail adaptive averaging," *J. Electron. Imaging*, vol. 17, no. 4, pp. 043005–, 2008.

[8] T. Chen, K. K. Ma, and L. H. Chen, "Tri-state median filter for image denoising," *IEEE Trans. Image Process.*, vol. 8, pp. 1834–1838, 1999.

[9] R. C. Hardie and C. G. Boncelet, "LUM filters: A class of rank-order-based filters for smoothing and sharpening," *IEEE Trans. Signal Process.*, vol. 41, pp. 1061–1076, 1993.

[10] S. J. Ko and Y. H. Lee, "Center weighted median filters and their applications to image enhancement," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 984–993, 1991.

[11] S. M. Guo, C. S. Lee, and C. Y. Hsu, "An intelligent image agent based on soft-computing techniques for color image processing," *Expert Systems With Appl.*, vol. 28, pp. 483–494, 2005.

[12] S. Schulte, V. De Witte, M. Nachtegael, D. Van der Weken, and E. E.Kerre, "Fuzzy random impulse noise reduction method," *Fuzzy Sets Syst.*, vol. 158, pp. 270–283, 2007.

[13] S. Schulte, M. Nachtegael, V. De Witte, D. Van der Weken, and E. E.Kerre, "A fuzzy impulse noise detection and reduction method," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1153–1162, 2006.

[14] F. Russo, "Fire operators for image processing," *Fuzzy Sets Syst.*, vol. 103, no. 2, pp. 265–275, 1999.

[15] F. Russo, "Hybrid neuro-fuzzy filter for impulse noise removal," *Pattern Recognit.*, vol. 32, pp. 1843–1855, 1999.

[16] H. K. Kwan, "Fuzzy filters for noise reduction in images," in *FuzzyFilters for Image Processing*, M. Nachtegael, D. Van der Weken, D. Van De Ville, and E. E. Kerre, Eds. Heidelberg, Germany: Springer, 2003, pp. 54–71.

[17] J. H.Wang, W. J. Liu, and L. D. Lin, "An histogram-based fuzzy filter for image restoration," *IEEE Trans. Syst. Man Cybern. B, Bern.*, vol. 32, no. 2, pp. 230–238, 2002.

◈◈◈

Code Convertor For Portability of Applications For ANDROID & iPHONE

Nitish Sharma, Swapnil Naik, Rasika Kulkarni & Tanvi Gokhale

University of Pune, India

Abstract - With all the latest technologies and techniques being implemented, Cell phones are no more used for just calling or messaging. They are at a stage where they can be used for doing almost anything and everything. The leading cell phones in today's tech-race are the Android phones and the iPhones of Apple. These two are leading the market when it comes to phones with latest and leading technology.

iPhone applications are developed in Objective C language while Android applications are developed in Java. Due to the current restrictions and differences between iPhone and Android platforms, applications that need to be deployed on both the platforms need to be developed twice. This involves double effort and time. Hence, there is a rise in demand, for Java to Objective C translator. This translator will allow an application to be developed only once but deploying it on both the platforms, i.e. iPhone and Android. Whenever any new application is to be developed, its application code along with the translation details are sent to the translator. If the application to be developed is for an iPhone, then the translator will refer to the Objective C library and generate a respective Objective C code for that application. If it is to be developed for an Android phone, then the translator will refer to the java library and generate java code for that same application. The generated code will then be sent to be implemented on the required platform. This would help in reducing the development time and energy.

Languages that will be used to implement this technique are Java, for the translator and Android, and Objective C for the iPhone.

I. INTRODUCTION

Whenever a developer has to create a new software application for smart phones, he will simply have to code it once, and using the FONEGAP, he can convert the application code so as to run it on both platforms, ios as well as Android.

II. DRAWBACKS OF THE CURRENT METHODOLOGY

The Traditional interfaces have some crossplatform related problems as :

1. The previous interfaces were not able to achieve all the native behavior of all the phone models.
2. There was a limitation that phone specific applications were not completely modularized to work on specific phone.
3. The same code can't be used for any other phone some modifications were required.

The new interface have a fairly good efficiency but the search for a more efficient method is still an area of exploration.

III. Objective of Proposed INTERFACE

This Translator will allows an application to be developed only once and deployed on both, iPhone & Android platforms, without any changes Simplify development and coding of applications to be deployed on the above mentioned platforms. Mobile platform specific applications can also be translated for multiple platforms.

IV. INPUT SPECIFICATION

Description and Priority

The system allows an application developer to create his application, and then via the utilities of the system convert them into the required code format to support any of the two required mobile platforms.
The system will also ask the developer to enter the conversion specification, on the basis of which, the final code format will be decided. The system will access the two language libraries, namely Java language library and Objective C library, as per required to generate the final code with respect to the conversion specification. The system converts the code and then dispatches it to the developer to be implemented on the required platform.

Stimulus/Response Sequences

User will provide the application code with the code conversion specifications. The system uses Java and Objective C librarys .

Functional Requirements

Application code written in Java Language.
Conversion details.

V. SYSTEM IMPLEMENTATION PLAN ::

The system mainly contains two components that are to be implemented. Firstly the convertor that converts the developed code into the required code format. Second, the two language libraries, namely the Java language library and the Objective c library, that help in code mapping.

The system allows an application developer to create his application, and then via the utilities of the system convert them into the required code format to support any of the two required mobile platforms.

The system will also ask the developer to enter the conversion specification, on the basis of which, the final code format will be decided. The system will then access the two language libraries, namely Java language library and Objective c library, as per required to generate the final code with respect to the conversion specification.

The system converts the code and then dispatches it to the developer to be implemented on the required platform.

VI. SOFTWARE QUALITY ATTRIBUTES :

The application has the ability to adapt for mobile phone that supports Android or iPhone. It also gives justice to the other important quality attributes such as

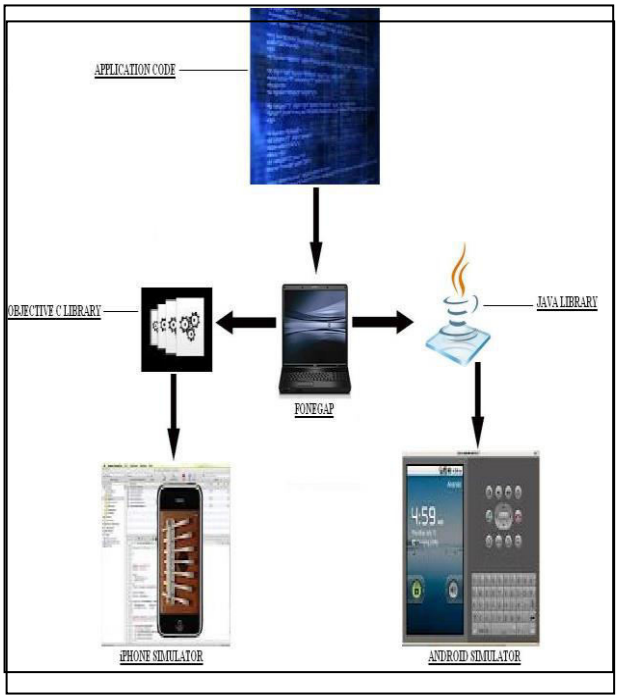
Correctness,

Flexibility,

Maintainability,

Robustness,

Reliability



Main System Architecture

VII. CONCLUSION AND FUTURE WORK

Can be extended for other quickly developing platforms like Blackberry. Mobile platform specific applications can also be translated for multiple latforms. Currently, a single call from the call log in an iPhone cannot be deleted like in other phones. Instead the whole call log has to be deleted. Thus, using our proposed interfacing technique this problem can be tackled.

VIII. APPLICATIONS

1. The Fonegap itself is an application software which helps acquire a code suitable for the two mentioned platforms.
2. It can also be applied to various other software systems which demand such conversions.
3. It can also be implemented to overcome the drawbacks of the two mentioned platforms with respect to their phone specific applications.

REFERENCES

[1] A compariosn of open and closed mobile platforms Hee-Yeon Cho; Choon-Sung Nam; Dong-Ryeol Shin; Sch. of Inf. & Commun. Eng., Sungkyunkwan Univ., Suwon, South Korea

[2] Android: Changing the Mobile Landscape. Butler, M.

[3] iPhone: Smarter Than the Average Phone Want, R.; Intel Labs., Santa Clara, CA, USA

[4] Open Android-For better and for worse [Tools & Toys] Proffitt, B.

Websites-

- www.developer.android.com
- www.androidapplication.com
- www.apple.com

◈◈◈

Privacy Preserved Centralized Model for Counter Terrorism

Abhishek Sachan[1] & Devshri Roy[2]
[1&2] Computer Science Maulana Azad National Institute of Technology, Bhopal, India

Abstract -Privacy preservation is an important aspect in field of counter terrorism. In the present scenario terrorist attacks are biggest problem for the mankind and whole world is under constant threat from these well-planned, sophisticated and coordinated terrorist operations. Now every country is focusing for counter terrorism. Government agencies are collecting the data from various sources and using that data to connect the dots to detect the terrorist group's activities and prevent the peoples from terrorist attacks. There are some chances that information may be misused by agencies. Different countries are having government agencies which are dealing with the counter terrorism but they are not sharing the data with each other because they don't want to disclose sensitive data. Alone a country can't fight against the terrorism. In this paper we are proposing a model so that these agencies can share the information without violation of the privacy.

Keywords - *privacy preservation; counter terrorism, data mining; surveillance.*

I. INTRODUCTION

Today, terrorist attacks are biggest problem for the mankind and whole world. Terrorists are those individuals who plan, participate in, and execute acts of terrorism. According to Brian Jenkins of the Rand Corporation terrorism is "the calculated use of violence such as fear, intimidation or coercion, or the threat of such violence to attain goals that are political, religious, or ideological in nature. Terrorism involves a criminal act that is often symbolic in nature and intended to influence an audience beyond the immediate victims."[1].

Counter-terrorism is the practices, tactics, strategies, and techniques that governments, militaries and police uses to prevent or in response to terrorist threats, both real and imputed. Counterterrorist operatives are engaged in the battle against terrorism. They may be agents of a state or country, including intelligence agents, investigators, and military personnel; or they may be law enforcement officers working at state or local levels. Some time private security and corporate security personnel may also be engaged in counterterrorism operations [1].

The term privacy is used frequently in ordinary language, until now there is no single definition of privacy [10]. The concept of privacy has broad historical roots in sociological and anthropological discussions about how extensively it is valued and preserved in various cultures [11]. Historical use of the term is not uniform, and there remains some confusion over the meaning, value and scope of the concept of privacy [13]. Privacy refers to the right of users to conceal their personal information and have some degree of control over the use of any personal information disclosed to others [12].

Security and privacy are related to each other we have to develop the system with privacy-protection technologies to protect civil liberties. Coordinated policies can help bind the two to their intended use [18]. Privacy-preserving is an important concern in the application of data mining techniques to datasets. Datasets contain personal, sensitive, or confidential information. Data distortion is a technique to preserve privacy in security-related data mining applications, such as in data mining-based terrorist analysis systems [2].

Today, data is one of the most important corporate assets of companies, governments, and research institutions [3] and is used for various private and public interest. The use of data mining technologies in counter terrorism and homeland security has been flourishing since the U.S. Government encouraged the use of information technologies [4]. Government access and use of personal information in commercial databases raise concerns about the protection of privacy and due process [5].

Data can be collected at a centralized location or collected at different locations, but integrated at a centralized location (data warehousing). Alternatively, data can be collected and stored at distributed locations. Different data storage patterns may have different privacy concerns. If the data storage is centralized, the major privacy concern is to shield the exact values of the attributes from the data analysts. Thus, data distortion is a technique that is usually considered in such a situation [6, 7]. On the other hand, in a distributed database situation, the major privacy concern is to maintain the independence of the distributed data ownership and to prevent the exchange of exact values of the attributes between different parties of the distributed database ownership.

This concern is related to the issue of data mining in a distributed environment [2, 8, 9].

It is necessary that data mining technologies designed for counterterrorism and security purpose have sufficient privacy awareness to protect the privacy of innocent people. Unfortunately, most existing data mining technologies are not very efficient in terms of privacy protections, as they were originally developed mainly for commercial applications, in which different organizations collect and own their databases, and mine their databases for specific commercial purposes. In the cases of security and counterterrorism, data mining may mean a totally different thing. Government may potentially have access to any databases and may extract any information from these databases. This potentially unlimited access to data and information raises the fear of possible abuse [2].

Telephone companies are sharing the telephone records of millions of peoples with the security agency. Security agency can use this information to create a database of detailed information for every telephone call made within the country. Intelligence agency then mined this database to uncover hidden terrorist networks [14].

People expect from their government to protect them from enemy attacks along with their civil liberties and privacy. Personal privacy is only violated if the violated party suffers some tangible loss, such as unwarranted arrest or detention. Privacy-protection technology is a key part of the solution not only to protect privacy but also to encourage the intelligence, law enforcement, and counterterrorism communities to share data without fear of compromising sources and methods [18].

Advanced information technologies offer key assets in confronting a secretive, asymmetric, and networked enemy. The policies must ensure that these powerful technologies are used responsibly and that privacy and civil liberties remain protected. People expect from their government to protect them from terrorist attacks, but fear the privacy implications of the government's use of powerful technology controlled by regulation and oversight. Some people believe the dual objectives of greater security and greater privacy present competing needs and require a trade-off; others disagree [15, 18, 19, 21].

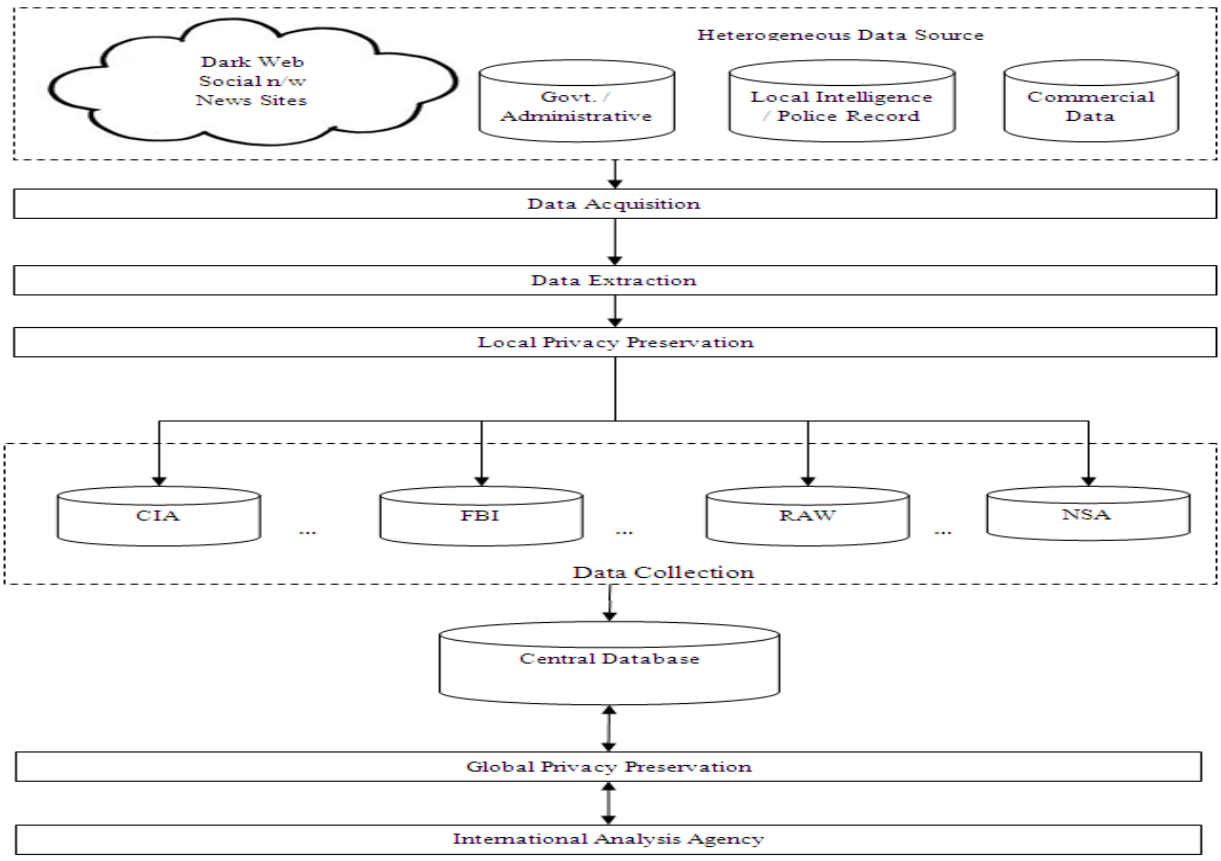


Figure 1. Privacy Preserved Centralized Model for Counter Terrorism

II. MODEL ARCHITECTURE

Privacy Preserved Centralized Model is proposed for Counter Terrorism. When we are discussing about counter terrorism one problem is coming that how can we allow our government agencies to perform surveillance over us. People believe that their privacy may be violated. Another problem is that different countries are having security agencies which are dealing with the counter terrorism but they are not sharing the data with another country because they don't want to disclose their sensitive data.

In this model we have tried to protect the privacy of individual along with the privacy of country's data so that data could be shared without violating privacy. Local privacy preservation module is use to preserve the privacy of the people's during surveillance & data collection. Government security agencies are increasingly moving towards data mining with the hope that advanced statistical techniques will connect the dots and uncover important patterns in large databases. Data surveillance technology is able to predict and prevent terrorist attacks, detect disease outbreaks, and allow for detailed social science research—all without the corresponding risks to personal privacy because machines, not people, perform the surveillance [14].

In this model central data mining concept is use, to solve the second problem of sharing the data between countries. Central database is the database in which data available all over the world is stored. International analysis agency is third party. No country can directly access this data. Agency will perform mining over this central data and return the desired data to the requested country. Even agency can't violet the privacy because it is restricted with the global privacy policies.

There are various privacy preservation techniques and technologies that can be applied over local and global privacy preservation modules. Privacy Appliance, Transformation Spaces, Immutable Audit, Selective Revelation, Self-Reporting Data, Anonymization and Inference Control are some privacy protection technologies [18]. Privacy preservation data mining techniques are k-anonymity, l-diverse, taxonomy tree, randomization, perturbation, condensation and cryptographic etc. that can be used based on the requirement [16, 17, 20, 22, 23].

III. CONCLUSION

We can say that by using this model security agencies can perform surveillance and data collection for counter terrorism without the violation of individual's privacy. This model helps to share data between agencies/countries without disclosure of country specific sensitive information. This model is uses both centralized and distributed data mining concept. Now the performance of the model is dependent on privacy preservation technique. If we will use strong/secure techniques in the model for privacy preservation then model will be strong/secure else weak.

REFERENCES

[1] Frank Bolz, Jr., Kenneth J. Dudonis, David P. Schulz, "The Counterterrorism Handbook Tactics, Procedures, and Techniques", In CRC Press, 2002.

[2] Shuting Xu, Jun Zhang, Dianwei Han, JieWang, "Singular value decomposition based data distortion strategy for privacy protection", In Knowledge and Information Systems, March 2006.

[3] Estvill-Castro V, Brankovic L, Dowe DL, Privacy in data mining. Australian Computer Society, NSW Branch, Australia. Available at w.acs.org.au/nsw/articles/ 1999082.html, 1999.

[4] Taipale KA, "Data mining and domestic security: connecting the dots to make sense of data", In Columbia Sci Tech Law Rev 5, 2003, pp. 1–83.

[5] Dempsey JX, Rosenzweig P, "Technologies that can protect privacy as information is shared to combat terrorism", Legal Memorandum #11, The Heritage Foundation. Available at www. heritage.org/ Research/ HomelandDefense/ lm11.cfm, 2004

[6] Agrawal D, Aggarwal CC, "The design and quantification of privacy preserving data mining algorithms", In Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems, Santa Barbara, California, USA, 2001.

[7] Liew CK, Choi UJ, Liew CJ, "A data distortion by probability distribution", In ACM Transaction Database System,1985, pp. 95–411.

[8] Agrawal R, Evfimievski A, Srikant R, "Information sharing across private databases", In Proceedings of the 2003 ACM SIGMOD international conference on management of data, San Diego, CA, 2003, pp. 86–97.

[9] Gilburd B, Schuster A, Wolff R, "K-TTP: a new privacy model for large-scale distributed environments", In Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining, Seattle, WA, USA, 2004.

[10] J.DeCew, privacy, The Stanford Encyclopedia of Philosophy, Editor:Edward N.Zalta, Summer 2002.

[11] A. F. Westin, The Right to Privacy, Atheneum 1967.

[12] S. Cockcroft and P. Clutterbuck, "Attitudes towards information privacy", In Proceedings of the 12th Australasian Conference on Information Systems, Australia, 2001.

[13] Justin Zhan, "Privacy Preserving Collaborative Data Mining", In IEEE, 2007.

[14] Simson L. Garfinkel, Michael D. Smith, "Data Surveillance" , In IEEE SECURITY & PRIVACY,2006,pp.15-17.

[15] R. Popp and J. Yen, eds., "Emergent Information Technologies and Enabling Policies for Counter-Terrorism", In Wiley & Sons/IEEE Press, 2006.

[16] PinkasB., "Cryptographic Techniques for Privacy-PreservingDataMining" In ACM SIGKDD Explorations, 4(2), 2002.

[17] S. Laur, H. Lipmaa, and T. Mielik¨ainen, "Cryptographically private support vector machines", In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 618–624.

[18] Robert Popp, John Poindexter, "Countering Terrorism through Information and Privacy Protection Technologies", IEEE SECURITY & PRIVACY, 2006, pp.18-27.

[19] Report to Congress Regarding the Terrorism Information Awareness Program, DARPA, May 2003; [response to Consolidated Appropriations Resolution, Pub. L. no.108-7, div. M, sec. 111(b), 2003].

[20] Charu C. Aggarwal and Philip S. Yu, "A condensation approach to privacy preserving data mining", In EDBT, 2004, pp. 183–199.

[21] J. Poindexter, "Overview of the Information Awareness Office," In DARPATech 2002, DARPA,2002, www.fas.org/irp/agency/dod/poindexter.html.

[22] Machanavajjhala A., Gehrke J., Kifer D., and Venkitasubramaniam M, "l-Diversity: Privacy Beyond k-Anonymity", In ICDE, 2006.

[23] E.Poovammal ,Dr. M. Ponnavaikko, "An Improved Method for Privacy Preserving Data Mining", In IEEE International Advance Computing Conference Patiala, India, 2009.

◈◈◈

Gesture Recognition based on Spatio-Temporal Trajectory in 3D Space Using Hidden Markov Models

Zeeshan Ali Sayyed, Sridhar Rajagopalan., Kiran Bhor, Raisa Naikwadi, Archana Shirke

Information Technology Department, Fr. C. Rodrigues Institute of Technology, Navi Mumbai, India

Abstract -This paper proposes a system for recognizing single hand gestures in three dimensional space on the basis of their spatio-temporal trajectories. The proposed system is based on Hidden Markov Model. Forward HMM topology with states ranging from 3 to 5 was used to model the gestures. The proposed system uses orientation of hand as the feature for training the Hidden Markov Models. The orientations were quantized and coded in order to shrink the observation space. A set of closely related gestures were trained and experiments were performed on them to decide various parameters (quantization factor and number of states in HMM) for optimum functioning of the system. This method lays foundation for sign language recognition.

Keywords - *Single Hand Gesture Recognition, Hidden Markov Model, Microsoft Kinect, Spatio-temporal.*

I. INTRODUCTION

Hand Gesture recognition that can contribute to natural man machine interface is still a challenging task. A gesture is a spatio-temporal pattern which may be static, dynamic or both. Static morphs of the hands are called postures and hand movements are called gestures [1]. In the past, various methods have been used to perform gesture recognition by Neural Networks [2], Hidden Markov Models [1][3], Fuzzy Systems [4], etc. HMM is a doubly stochastic statistical model[11] and is capable of modeling spatio-temporal time series where the same gesture can differ in shape and duration. Recognition becomes increasingly difficult when the observation space increases. Hence, various clustering and quantization methods are used to shrink it.

This paper describes a quantization and coding based method employed for recognizing 3-D gestures on the basis of their spatio-temporal trajectory using Hidden Markov Models. It describes the way data was collected, processed and finally used for training the system. Baum-Welch algorithm[5] was used for training whereas forward - backward Algorithm[9][10] was used for recognition. Actual gestures were trained and the results were analyzed. This paper is organized as follows: The experimental design and setup are described in Section 2 whereas the Results and Analysis are elaborated in Section 3. Finally, Section 4 consists of Conclusion and Future plans.

II. EXPERIMENTAL SETUP AND DESIGN

The various aspects of the experiment are described in the following subsections:

A) Data Collection

The experiment was aimed at recognizing gestures in three dimensional space. Such three dimensional data can be obtained in various ways like using a data glove [2], using multiple cameras or using a depth camera [6]. Microsoft Kinect which works on the principle of depth camera was selected for this purpose as it is cheaper than data glove and more robust compared to use of multiple cameras.

Microsoft Kinect captures and returns the position of as many as 20 joints in the human body. This position includes the X, Y and Z co-ordinates of the joints in three dimensional space at a particular instant of time. Also, Kinect captures 30 frames per second on an average. Kinect SDK version 1 Beta 2 was used as an interface for interacting with Kinect.

Four similar gestures were selected for the purpose of this experiment as shown in Fig. 1. A total of 150 gesture samples for each gesture were collected from 10 different subjects with different physical attributes to provide for variety in samples. Slight variations were incorporated in the sample in terms of position of subject, speed of gesture, style of gesture etc. to make the data wide.

B. Feature Selection

Choosing appropriate features for training from the collected gesture data is a critical aspect of any Gesture Recognition system. Three features can be considered for this purpose viz. position velocity and orientation. From [7], it can be noted that orientation yields the best results. Hence, it was selected for the purpose of this experiment.

Orientation represents the direction of motion. It can be represented in terms of angles in XY, YZ and XZ planes (α, β and γ respectively). α, β and γ are constrained by the following relation:

$$\cos^2 \alpha + \cos^2 \beta + \cos^2 \gamma = 1$$

Hence only α and β were used to eliminate redundancy. They are calculated as follows:

$$\alpha = \tan^{-1} \frac{y2-y1}{x2-x1} \quad \beta = \tan^{-1} \frac{z2-z1}{y2-y1}$$

where (x1, y1, z1) and (x2, y2, z2) represent the positions at time (t-1) and t respectively.

C) Quantization

In the above feature selection method, the possible number of observations is 12960 which is quite large. Thus, it is crucial to reduce it to bring down the variance of the system. In this technique the observation range of α and β is divided into n partitions and all values in a given partition are represented by its partition numbers p and q respectively, where

$$p, q \in \{x \mid x \text{ is an integer and } 0 \leq x < n\}$$

Now the values p and q are represented by a unique code(c) calculated as follows:

$$c = n \times p + q$$

where c takes values from 0 to n^2. Thus the observation space is significantly reduced.

D) Training and Testing

Baum Welch algorithm [5] was used for training the system. Every gesture is represented by a unique HMM and thus has its own set of parameters $\lambda = (\pi, A, B)$ where

π = Set of initiation parameters.
A = State Transition matrix
B = Observation Matrix

A given gesture or an HMM consists of multiple states. The states are hidden and define the complexity of the function represented by the model. Decreasing the number of states results in selection of function which is not complex enough to represent the data and increasing it results in a function which over-fits the data. In order to find the optimum number of states, models with 3, 4 and 5 states were trained and tested. Generally the states are interconnected in such a way that any state can be reached from any other state (ergodic model). However a variation called the forward model (or left-right model) is found to suit temporal problems better. In this model only forward state transitions are allowed [8] i.e.

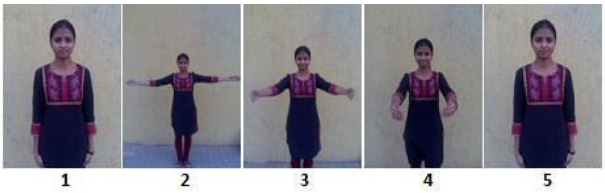
$$A[i, j] = 0 \text{ for all } i > j$$

As gestures have a temporal component we have decided to use the forward model.

Selecting the quantization parameter n also affects the size of the observation space and hence has to be selected carefully to maximize the accuracy of the system. Hence models using n = 8, 9, 10 and 12 were trained and tested.

To choose the most optimum number of states and quantization parameter, the experiment was repeated 5 times. Each time, the set of all samples was randomly partitioned into a training set of 135 samples and testing set of 15 samples. The final results were computed by averaging the results of each experiment.

Gesture 1



Gesture 2

Gesture 3

Gesture 4

Fig. 1 : Trained Gestures

III. RESULTS AND ANALYSIS

The combination of states and quantization parameter which gives better overall gesture recognition accuracy along with lesser standard deviation between various gestures should be chosen. The latter is required because a good model is one which predicts all gestures equally well. A high standard deviation between gestures indicates bias of the system towards one or two particular gestures. To help us decide the right combination of parameters, the harmonic mean(M) of error and standard deviation was chosen:

$$\text{Harmonic Mean (M)} = \frac{2 \times error^2 \times \sigma}{error^2 + \sigma}$$

where error = 1 – accuracy
σ = standard deviation in error

The combination with the lowest value of M is chosen.

Tables I to IV give the Accuracy, Standard Deviations in error (σ) and Harmonic Means (M) of experiments performed for different no of states used in building HMMs for each gesture.

TABLE I. 3 STATES HMM

QF	Gesture ID(Accuracy)				System Accuracy	σ	M
	1	*2*	*3*	*4*			
8	0.87	0.9	0.84	0.9	0.87852	0.027	0.019
9	0.94	0.8	0.94	0.928	0.9035	0.069	0.016
10	0.88	0.85	0.85	0.924	0.88117	0.032	0.019
12	0.92	0.82	0.88	0.871	0.87852	0.041	0.021
Avera	**0.90**	**0.84**	**0.88**	**0.906**			

TABLE II. 4 STATES HMM

QF	Gesture ID(Accuracy)				System Accuracy	σ	M
	1	**2**	**3**	**4**			
8	0.871	0.84	0.65	0.9	0.817825	0.11	0.05
9	0.828	0.84	0.9	0.97	0.885675	0.06	0.02
10	0.885	0.71	0.95	1	0.88925	0.12	0.02
12	0.714	0.85	0.71	0.85	0.785675	0.08	0.05
Avg	**0.824**	**0.81**	**0.80**	**0.93**			

TABLE III. 5 STATES HMM

QF	Gesture ID(Accuracy)				System Accuracy	σ	M
	1	**2**	**3**	**4**			
8	0.942	0.81	0.87	0.85	0.8708	0.056	0.02
9	0.814	0.87	0.85	0.87	0.8542	0.027	0.02
10	0.825	0.82	0.82	0.95	0.85995	0.064	0.03
12	0.714	0.85	0.71	0.78	0.767825	0.068	0.06
Averag	**0.824**	**0.84**	**0.81**	**0.86**			

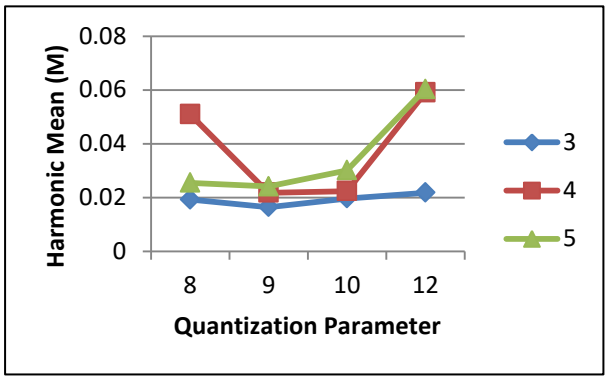


Fig. 2 : Harmonic Mean for different combinations of No of states and Quantization Parameter for 3, 4, 5 states.

From tables I, II, III and Figure 2 it is clear that a combination of 3 states and quantization parameter of 9 gives the lowest value of M. But it can also be seen that Gesture 4 gives better accuracy for 4 states while all other gestures give maximum accuracy for 3 states. Hence an experiment was carried out where Gesture 4 was allotted 4 states while all other gestures were confined to 3 states. The results of this experiment are shown in table 4.

TABLE IV. 3-3-3-4 STATES HMM

QF	Gesture ID(Accuracy)				System Accuracy	σ	M
	1	**2**	**3**	**4**			
8	0.9571	0.9285	0.514	1	0.84995	0.2257	0.041
9	0.8428	0.7857	0.857	0.9571	0.860675	0.0713	0.031
10	0.9428	0.9142	0.842	0.9571	0.914225	0.0508	0.012
12	0.8571	0.8285	0.757	0.8571	0.82495	0.0471	0.037
Averag	**0.8999**	**0.864**	**0.742**	**0.942**			

From Table IV it is clear that the combination of 3 states for Gesture 1, 2 & 3 and 4 states for Gesture 4 along with a quantization parameter of 10 gives the least value of M, hence was chosen as the final set of parameters. The accuracy for the chosen set of parameters is shown in the Fig 3.

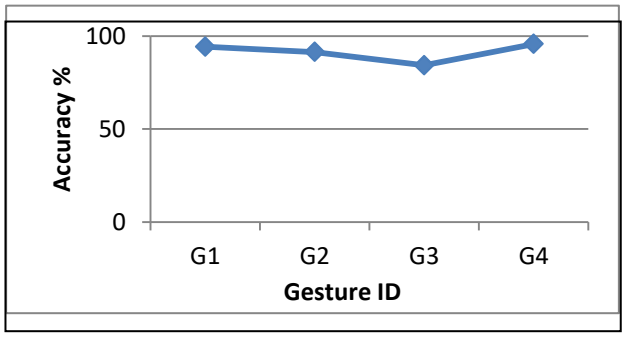


Fig. 3 : Accuracy for chosen Parameters

IV. CONCLUSION AND FUTURE WORK

The mean accuracy of the system for the chosen set of parameters is 91.42%. The standard deviation in error of various gestures is 0.0508. This is good enough for use in actual applications.

We plan to scale the system to enable it recognize more gestures. We also plan to use this work to implement a system capable of handling multi point gestures. The system can then be used to make a Sign Language Interpreter based on Indian Sign Language to assist the vocally and hearing impaired in India, which is the main motive behind this work.

REFERENCES

[1] A Hidden Markov Model-Based Continuous Gesture Recognition System for Hand Motion Trajectory by Mahmoud Elmezain et al. Otto-von-Guericke-University Magdeburg, Germany.

[2] Aleem Khalid Alvi, M. Yousuf Bin Azhar, Mehmood Usman, Suleman Mumtaz, Sameer Rafiq, Razi Ur Rehman, Israr Ahmed; International Journal of Information Technology 1(1) 2004 1-12 ; Pakistan Sign Language Recognition Using Statistical Template Matching;

[3] Hyeon-Kyu Lee; Kim, J.H. : Pattern Analysis and Machine Intelligence, IEEE Transactions: Oct 1999.

[4] E. Holden, R. Owens, and G. Roy. Hand Movement Classification Using Adaptive Fuzzy Expert System. *Expert Systems Journal, Vol. 9(4), pp. 465-480, 1996.*

[5] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Statist., vol. 41, no. 1, pp. 164–171, 1970.

[6] Xia Liu; Fujimura, K.; Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference 17-19 May 2004 (pg 529 – 534);

[7] M. Elmezain, A. Al-Hamadi, and B. Michaelis. RealTime Capable System for Hand Gesture Recognition Using Hidden Markov Models in Stereo Color Image Sequences. W S C G Journal, Vol. 16(1), pp. 65-72, 2008.

[8] A tutorial on Hidden Markov Model and selected applications in speech recognition, Rabiner.L.R, Proceedings of IEEE, Feb 1989.

[9] L.E. Baum and J.A. Egon, “An equality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology,” Bull. Amer. Meteorol. Soc., vol. 73, pp. 360-363, 1967

[10] L. E. Baum and G. R. Sell, “Growth functions for transformations on manifolds,” Pac. J. Math., vol. 27, no. 2, pp. 211-227, 1968.

[11] Hidden Markov Models for Gesture Recognition by Donald O. Tanguay, Jr.

◈◈◈

Energy-Efficient MAC Protocol for Wireless Sensor Networks - A Review

Smriti Joshi & Anant Kr. Jayswal
Amity University, Noida.

Abstract -Energy efficiency is the kernel issue in the designing of wireless sensor network(WSN) MAC protocols. Energy efficiency is a major consideration while designing wireless sensor network nodes. Most sensor network applications require energy autonomy for the complete lifetime of the node, which may span up to several years. These energy constraints require that the system be built such that Wireless sensor networks use battery-operated computing and sensing devices. A network of these devices will collaborate for a common application such as environmental monitoring. Each component consumes minimum possible power, ensure the average successful transmission rate, decrease the data packet average waiting time, and reduce the average energy consumption. Influencing by the design principles of traditional layered protocol stack, current MAC protocol designing for wireless sensor networks (WSN) seldom takes load balance into consideration, which greatly restricts WSN lifetime. As a novel Forwarding Election-based MAC protocol, is presented to prolong WSN lifetime by means of improving energy efficiency and enhancing load balance.

I. INTRODUCTION

A wireless sensor network (WSN) is a wireless network consisting of spatially distributed autonomous devices that use sensors to monitor physical or environmental conditions. These autonomous devices, or nodes, combine with routers and a gateway to create a typical WSN system. Each node consists of processing capability (one or more microcontrollers, CPUs or DSP chips), may contain multiple types of memory (program, data and flash memories), have a RF transceiver (usually with a single omni-directional antenna), have a power source (e.g., batteries and solar cells), and accommodate various sensors and actuators. The nodes communicate wirelessly and often self-organize after being deployed in an ad hoc fashion. Systems of 1000s or even 10,000 nodes are anticipated. Such systems can revolutionize the way we live and work.

A WSN system is ideal for an application like environmental monitoring in which the requirements mandate a long-term deployed solution to acquire water, soil, or climate measurements. For utilities such as the electricity grid, streetlights, and water municipals, wireless sensors offer a lower-cost method for collecting system health data to reduce energy usage and better manage resources. In structural health monitoring, you can use wireless sensors to effectively monitor highways, bridges, and tunnels. You also can deploy these systems to continually monitor office buildings, hospitals, airports, factories, power plants, or production facilities.

A. MAC Layer Protocol:

In a wireless sensor network the MAC Layer protocols are supposed to perform the following tasks:

1. To create an infrastructure and establish link for data transfer.
2. To share network communication resources between sensor nodes.

The MAC layer is responsible for access to shared medium. It assists nodes to decide when to access the shared medium. The MAC Protocols [1] can be classified as follows:

1. Scheduled:

This is based on Time Division Multiple Access (TDMA) protocol. In this mechanism the channel is divided into fixed time slots. A complete cycle of these slots is called frame. TDMA Protocols are inherently energy conserving as they reduce wastage due to Collision, Idle Listening and Overhearing.

2. Random:

This is based on Carrier Sense Multiple Access (CSMA) protocol. In random access protocols, the channel is allocated to nodes on demand. i.e. the nodes contend for channel and if they find the channel free, starts transmission else postpone the transmission until the channel becomes idle and sense the channel to grab the chance to transmit the channel.

II. WSN PROTOCOLS:

1) LEACH: LEACH

LEACH (Low Energy Adaptive Clustering Hierarchy) was the first sub-cluster-style routing protocols in WSN. LEACH conserves energy because it uses data compression techniques and sub-cluster dynamic routing technology. In LEACH algorithm each round is consist of two states:

(a) Setup State: during this step a cluster head is selected for that round.

(b) Steady State: In this phase the nodes send data to the cluster head.

The cluster head node is selected at random; it balances the network load and also the rapid death of cluster nodes. The probability of a node being the cluster node is given by:-

$$T(n) = \begin{cases} \dfrac{p}{1 - p * (r \bmod \frac{1}{p})} & if \quad n \in G \\ 0 & otherwise \end{cases}$$

Where P is the percentage of all cluster head, r is the number of current selection rounds, r mod (l/P) is selected the number of cluster head nodes in a cycle, G is not selected cluster head node set.

The deficiencies of LEACH are:-

The probability of a node which was once selected as a cluster head for again becoming a cluster head node becomes 1/p in the next recycling round. Also the initial energy of each node is assumed to be equal and energy consumption is equal when a node is selected as a cluster head. The energy cost is also increased as the algorithm frequently changes the cluster head.

IMPROVED LEACH ALGORITHM:

$$T(n)_{new} = \frac{p}{1 - p \times [r \bmod (1/p)]} \frac{E_{n\_current}}{E_{n\_\max}}$$

En-current is current energy of the node, *En-max* is initial energy. Formula is improved so that a lower proportion of energy consumption is selected as cluster head node priority.

This shows that as compared with leach, the proposed protocol considers the influence to route mechanism from the route hop number, node position, and energy consumption of each node.

Furthermore, it can reduce the delay of data forwarding and satisfy the demand of WSN application.

2) MULTI PARENT

Multi-Parent method has the advantage of reducing the delay the data will experience, while at the same time it uses an awake scheduling to reduce energy consumption. On the other hand, determining the parents for any node in the network is an algorithm that requires many processing cycles.

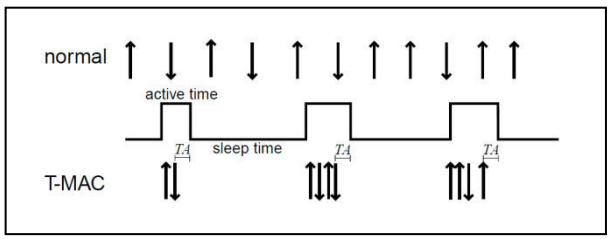
3) P_MAC (PENDULUM-MAC)

In this, the system is organized into layers and in each layer the nodes are assigned time slots in which they can be awake to report the data to the BS. Sensor nodes scattered in the geographic area for monitoring follow an awake scheduling. That means, the sensor nodes will determine when to wake up and when to go to sleep based on the current location of the node. But a major disadvantage lies with this technique that is the delay in collecting data for extended periods of time. Also it needs a certain density to operate that is, it cannot work on 250 node, it needs a denser network to work on so that a connected network is created. Thus it is not suitable for a small network.

4) TMAC(Timeout-MAC)

In T-MAC all the messages are transmitted in a burst of variable length and there is gap between the bursts called sleep/sleep time. This is to reduce the idle listening. The node awakes periodically to communicate with neighbors and it uses RTS and CTS, Data Acknowledgement (ACK) scheme, which provides both collision avoidance and reliable trans-mission.

In this the messages are stored in a buffer and then a frame is made to transmit containing messages during the active time as shown in fig. The active time ends when there is no active event for a time period TA and the node goes to sleep mode. At the time of high load nodes communicates continuously without sleeping.



The major disadvantage with this technique is “The early sleep problem”. i.e. the node goes to sleep mode even if it its neighboring node have something to send to it.

It has been found from previous research papers that T-MAC is more efficient than the traditional protocols, Pendulum and Leach protocol.

III. MAJOR ISSUES OF ENERGY WASTAGE:

1. Idle listening

When nodes have nothing to send or receive, the nodes still remain in active state and do idle listening to the network. This process consumes equal amount of energy as during transmitting or receiving process. Thus resulting into wastage of energy.

2. Collision or Corruption

Normally collision may occur when neighbouring nodes contend for free medium and lossy channel will result in corruption of transmitted packets. When either of two cases happens corrupted packets should be retransmitted, which increases energy consumption.

3. Overhearing

which happens when a node receives some packets that are destined to other nodes.

4. Control Packet Overhead

Exchanging control packets between sender and receiver also consumes some energy.

5. S-MAC(Sensor-MAC):

Sensor MAC (S-MAC) [3] is a contention based protocol specifically designed for wireless sensor networks. Its basic principle is CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance).It introduced a periodic "Listen and Sleep" method to avoid idle listening & to reduce the energy wastage. Each node follows a periodic sleep and listen schedule as shown in fig. In listen period, the node senses the network, if found idle, the node performs listening and communicate with other nodes. When sleep period comes, the node will try to sleep by turning off their radios. This significantly reduces the time spent on idle listening. In this protocol the nodes use the RTS (Ready to send), CTS (Clear to send) and Data Acknowledgement (ACK) to communicate. When a node finds a RTS or CTS packet destined for some other node, it goes to sleep mode. This is a periodic process. At the end of sleep mode the node wakes-up and look for some event, if not found it again go to sleep mode. S-MAC proposes a low-duty-cycle operation which reduces energy consumption.

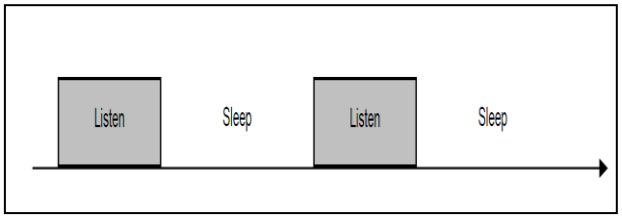


Fig. Periodic listen and sleep

A complete cycle of listen and sleep period is called a frame. During sleep period, the node will turn off its radio if possible. In this way, a large amount of energy consumption caused by unnecessary idle listen can be avoided especially when traffic load is light. The nodes in the network make a virtual cluster with its neighboring node and share a synchronization schedule for listen and sleep period. Thus there may exist more than one cluster in a network. In different clusters the nodes use periodic SYNC packet to find its neighbor. This process is called PND (Periodic Neighbor Discovery).

The S-MAC protocol uses the following to reduce or avoid the four major issues of energy wastage discussed above:

- The scheme of periodic listen and sleep reduces energy consumption by avoiding idle listening.
- The overhearing problem is avoided by using the in-channel signalling to put each node to sleep when its neighbour is communicating to another node.
- A complete synchronization mechanism, including periodic SYNC packets broadcast is used to avoid collision.
- S-MAC uses only a pair of RTS/CTS for one message passing but requests an ACK for each fragment. This reduces the control packet overhead to a great extant.

The S-MAC protocol essentially trades used energy for throughput and latency. Throughput is reduced because only the active part of the frame is used for communication. Latency increases because a message-generating event may occur during sleep time.

IV. CONCLUSION

This paper gives the performance analysis of all the protocols that have been proposed for wireless sensor networks till date. As compared to Leach and the Pendulum techniques MAC techniques are considered to be better as far as efficiency and performance of wireless sensor networks is concerned. Also the MAC.

Protocols have an interesting property that it has the ability to make trade-offs between energy and latency according to traffic conditions. But the problem of early sleep is observed in T-MAC and energy wastage issues are being observed in S-MAC. So further work will include the analysis of these issues.

REFERENCE

1. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "Wireless Sensor Networks: A Survey", Elsevier Science B.V., pp.393-422, Dec.2002.
2. Wei Ye, John Heidemann, Deborah Estrin "An Energy-Efficient MAC Protocol for Wireless Sensor Networks", INFOCOM 2002. Twenty-First Annual Joint Conferences of the IEEE Computer and Communications Societies. Proceedings. IEEE
3. Holger Karl and Andreas Willig, "Protocols and Architectures for wireless sensor networks", John Wiley & Sons Ltd, 2005.
4. Qingchun Ren and Qilian Liang "An Energy-Efficient MAC Protocol for Wireless Sensor Networks", Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE
5. Sung-Chan Choi; Jang-Won Lee; Yeonsoo Kim; Hakjin Chong; "An energy-efficient mac protocol with random listen-sleep schedule for wireless sensor networks", TENCON 2007 - 2007 IEEE Region 10 Conference
6. Giuseppe Anastasi, Marco Conti, Mario Di Francesco and Andrea Passarella, "Energy conservation in wireless sensor networks: A survey", Elsevier B.V., vol.7, pp.537-568, July 2008
7. Hongbin Chen "The Role of Recharging in Energy Efficiency for Wireless Sensor Networks", Wireless Communications and Signal Processing (WCSP), 2010 International Conference
8. Hung-Chi Chua et al.; Ying-Hsiang Liao; Lin-Huang Chang; Fang-Lin Chao "A Level-based Energy Efficiency Clustering Approach for Wireless Sensor Networks", Ubiquitous, Autonomic and Trusted Computing, 2009. UIC-ATC '09. Symposia and Workshops
9. Haithem Ben Chikha, Amira Makhlouf and Wiem Ghazel "Performance Analysis of AODV and DSR Routing Protocols for IEEE 802.15.4/ZigBee", Communications, Computing and Control Applications (CCCA), 2011 International Conference
10. Murizah Kassim, Ruhani Ab. Rahman, Roihan Mustapha "Mobile Ad Hoc Network (MANET) Routing Protocols Comparison for Wireless Sensor Network", System Engineering and Technology (ICSET), 2011 IEEE International Conference
11. Wei Ye, John Heidemann, Deborah Estrin, "An Energy-Efficient MAC Protocol for Wireless Sensor Networks" INFOCOM 2002. Twenty-First Annual Joint Conferences of the IEEE Computer and Communications Societies. Proceedings. IEEE
12. Ma, C.; Ma, M.; Yuanyuan yang, "Data-centric energy efficient scheduling for densely deployed sensor networks" Dept. of Comput. Sci., State Univ. of New York, Stony Brook, NY, USA Communications, 2004 IEEE International Conference
13. Liang Zhao; Xiang Hong; Qilian Liang,"Energy-efficient self-organization for wireless sensor networks: a fully distributed approach",Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE
14. Boscardin, T.; Songlin Cai; Gao, R.X.; Weibo Gong,"Energy efficient MAC protocol for condition monitoring sensor networks" Decision and Control, 2004. CDC. 43rd IEEE Conference on
15. Horton, M.; Suh, J.,"A vision for wireless sensor networks", Microwave Symposium Digest, 2005 IEEE MTT-S International
16. Qingchun Ren; Qilian Liang,"An energy-efficient MAC protocol for wireless sensor networks", Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE
17. Tabassum, N.; Urano, Y.; Ahsanul Haque, A.K.M, "GSEN: An Efficient Energy Consumption Routing Scheme for Wireless Sensor Network", Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, 2006. ICN/ICONS/MCL 2006. International Conference on
18. Sung-Chan Choi; Jang-Won Lee; Yeonsoo Kim; Hakjin Chong, "An energy-efficient mac protocol with random listen-sleep schedule for wireless sensor networks", TENCON 2007 - 2007 IEEE Region 10 Conference
19. Tao Zhang; Lijun Chen; Daoxu Chen; Li Xie, "EEFF: A Cross-Layer Designed Energy Efficient Fast Forwarding Protocol for Wireless

Sensor Networks", Wireless Communications and Networking Conference, 2009. WCNC 2009. IEEE

20. Alsaify, B.A.; Thompson, D.R.,"Pendulum: An energy efficient protocol for Wireless Sensor Networks" Sensors Applications Symposium (SAS), 2010 IEEE
21. Lejiang Guo; Weijiang Wang; Jian Cui; Lan Gao,"A Cluster-Based Algorithm for Energy-Efficient Routing in Wireless Sensor Networks", Information Technology and Applications (IFITA), 2010 International Forum
22. Mankar, G.; Bodkhe, S.T.,"Traffic aware energy efficient routing protocol", Electronics Computer Technology (ICECT), 2011 3rd International Conference
23. Seokjin Sung; Hyunduk Kang; Eunchan Kim; Kiseon Kim, "Energy Consumptlon Analysls of S-MAC Protocol in Single-Hop Wireless Sensor Networks", Communications, 2006. APCC '06. Asia-Pacific Conference on
24. Zohaib, M.; Jadoon, T.M., "Comparison of S-MAC & TDMA-W Protocols for Energy Efficient Wireless Sensor Networks", Emerging Technologies, 2006. ICET '06. International Conference
25. Jiangtao Wang; Geng Yang; Shengshou Chen; Yanfei Sun, "Secure LEACH routing protocol based on low-power cluster-head selection algorithm for wireless sensor networks", Intelligent Signal Processing and Communication Systems, 2007. ISPACS 2007. International Symposium
26. Wei Bo; Hu Han-ying; Fu Wen, "An Improved LEACH Protocol for Data Gathering and Aggregation in Wireless Sensor Networks", Computer and Electrical Engineering, 2008. ICCEE 2008. International Conference
27. Woonsik Lee; Minh Nguyen; Verma, A.; Hwang Lee, "Schedule unifying algorithm extending network lifetime in S-MAC-based wireless sensor networks", Wireless Communications, IEEE Transactions
28. Haifang Feng; Lixiang Ma; Supeng Leng,"A low overhead wireless sensor networks MAC protocol", Computer Engineering and Technology (ICCET), 2010 2nd International Conference
29. Abo, R.; Barkaoui, K.; Djouani, K., "Verification and Performance Evaluation of S-MAC Protocol Based on Process Calculi", Distributed Computing Systems Workshops (ICDCSW), 2010 IEEE 30th International Conference.
30. Zhang, Yu-quan; Wei, Lei, "Improving the LEACH protocol for wireless sensor networks", Wireless Sensor Network, 2010. IET-WSN. IET International Conference

◈◈◈

Novel Techniques to Eradicate Energy Inefficiencies That Abbreviate The Lifetime of The Cell Phone Based WSNs

Wilson Thomas[1] & Lajish V. L[2]

[1] Department of Computer Science , JJTU Jhunjhunu, Iindia
[2] Department of Computer Science Calicut University, Kerala,India

Abstract **-**The Cell Phone Based WSN of compressed micro-sensors for data acquirement and supervise some surroundings distinctiveness, such as noise, trembling, temperature, and strain. These sensors are entrenched devices accomplished of data communication. In numerous of applications, sensor nodes are deployed over a geo-graphically large region. Due to their configuration, data of measured values must be transferred among stations through these sensor nodes. For this reason a successful, energy efficient routing protocol should be implemented to avoid data loss and additional challenges within limited energy levels. This paper presents a cell phone based routing algorithm for wireless sensor networks, based on the selection of the scheme of dynamic nodes. The key objective is to boost the lifetime of a sensor network while not cooperation data delivery. Significant tasks such as, scrutinize, supervise and determine of energy levels of nodes are handled by these independent mechanisms.

Keywords - *Abbreviate of Lifetime WSNs protocol, Cell Phone, Cluster-based Routing Protocol, Mobility Factor etc.*

I. INTRODUCTION

The speedy enlargement of technology has specified increase to a novel class of distributed systems known as Cell Phone Based Wireless Sensor Networks. A WSN consists of several sensor nodes that have the capability to converse among themselves using radio antenna. These nodes are tiny in size with limited memory, energy source and processing power. Hence they all work together in collaboration as a network towards reaching a general goal of sensing a physical parameter over a large geographic area with superior accuracy. The Wireless Sensor Networks are measured as influential sensing network to the present day world due to their agreeable to support a diversity of real-world applications. The elasticity in its use is also the cause for it to be a demanding research and engineering problem. Wireless Sensor nodes are constrained in energy provide and bandwidth. Thus, inventive techniques that eradicate energy inefficiencies that would abbreviate the lifetime of the network are extremely required. Such constraints collective with a distinctive deployment of big number of sensor nodes pretense many challenges to the supervision of Wireless Sensor Networks and require energy consciousness at all layers of the networking protocol stack. Several new algorithms have been proposed for the routing problem in Wireless Sensor Networks. These routing mechanisms have taken into consideration the inherent features of WSNs along with the application and architecture requirements. The task of finding and maintaining routes in WSNs is nontrivial since energy restrictions and sudden changes in node status (e.g., failure) cause frequent and unpredictable topological changes. Routing techniques proposed here employ some well-known routing tactics to minimize latency and energy consumption e.g., data aggregation and clustering.

II. RELATED WORK

Routing in Wireless Sensor Networks is extremely demanding due to the intrinsic characteristics.

- Since the addressing scheme is not well appropriate. Enormous number of nodes makes it more complex. Thus addressing scheme cannot be solved by conventional IP based protocols.
- Sensor networks necessitate supervision for transferred data approximately all characteristic applications of communication networks.
- Sensor nodes are usually cell phone based and difficult to determine location on geographical area. Global Positioning System provides some sort of information but it's not a practicable solution.
- Many new algorithms have been proposed for the routing problem in WSNs due to dissimilar scenarios and dissimilar situations. None of them overcome above challenged.

In common, routing in WSNs can be divided into three main categories such as data-centric routing, hierarchical based (cluster based) routing, and location based routing depending on the network structure. In flat based routing all nodes plays the same role and it is not feasible to assign a global identifier to them. Base Stations sends queries and waits for data from the

sensors. Well known protocols proposed are the Sensor Protocol for Information via Negotiation [7], [8], Directed Diffusion [9], Rumor Routing [10], Minimum Cost Forwarding Algorithm [11], Gradient based Routing [12], Information driven sensor Querying [13] In a hierarchical architecture, sensor nodes are grouped and the one with the greatest residual energy is usually chosen as the cluster head. Higher energy nodes can be used to process and send the information, while low energy nodes can be used to perform the sensing task of the environment. This routing also called cluster based routing method. Some of the proposed cluster based protocols are the Low-Energy Adaptive Clustering Hierarchy (LEACH) [13], Power-Efficient Gathering in Sensor Information Systems (PEGASIS) [14], Threshold sensitive Energy Efficient sensor Network protocol (TEEN) [15], the location information of the sensor nodes is elegantly utilized in order to determine energy efficient routing paths. The distance can be estimated according to the level of signal strength. To save energy, some location based schemes demand that nodes should go to sleep if there is no activity. Well known protocols in this category are the Minimum Energy Communication Network (MECN) [4]. Geographic Adaptive Fidelity (GAF) , Geographic and Energy Aware Routing (GEAR) , Most Forward within Radius (MFR) etc.

III. PROPOSED CELL PHONE BASED ROUTING PROTOCOL

We present the effective standard of the proposed Cluster Protocol for Cell Phone Based WSN with Pseudo -code. The Abbreviate of Lifetime WSNs protocol (ALWP) works into the following phases.

A. Cluster start assortment

The Clusters are formed based on the environmental locations of sensors by base station and selects cluster heads based on the remaining energy and position of the sensors. Since all nodes, have the same preliminary energy cluster heads, is selected based on a random number between zero and one and cluster heads probability, which is comparable to the method used in the Low-Energy Adaptive Clustering Hierarchy protocol [2, 4]. Once cluster heads are selected they broadcast their positions and identification details. A node N is assigned to a cluster if the cluster heads of that cluster is at the minimum Distance with N. The node N then sends a registration message to the cluster heads with its identification details and current position. Cluster heads send cluster information to Base Station for centralized control and operations. We assume that each cluster heads that is selected at the beginning of a round is static until a new cluster heads is selected in the next round based on the mobility factor of nodes. After a number of rounds a new cluster formation and cluster heads selection phase is initiated to balance the network energy consumptions. Once the network process starts and nodes move at a fixed speed, each node keeps track of the number of movements inside and outside of its recent cluster based on which node's mobility is calculated at each round.

Cluster start assortment Pseudo -code:

```
struct node
{
char[10] clusterarea;
int numberofclusters;
int clusterid;
struct node *next;
}*p;
void initial state(int clusterformation,char*
initialCHselection)
int noofnodes;
void main()
{
int i;
char[10] k;
intial state(i,&k);
}
initial node(clusterinformation,initialCHselection)
{ for (int i=1;i<=noofnodes;i++)
{ for(int j=1;j<=noofcluster;j++)
{ for(int k=1;k<=i;k++)
{ if(position [node[i]]==clusterarea[clusterid[j]]
{ node[i]<-clusterid[j]}
}}} for(j=1;j<=i;j++)
{ CHprobnode[i][j]<-random(0,1)
if (CHprobnode[i][j] < CHprobability)
{ CH[i] <-node[j]; }
  }
```

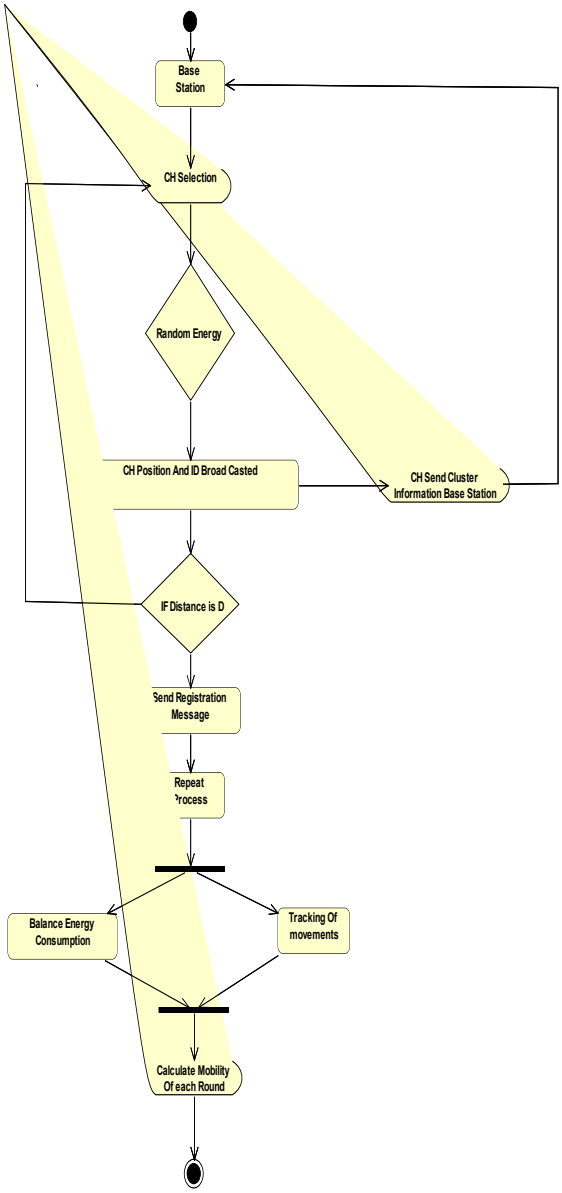


Figure : 1

B. Unwavering segment

In the *Unwavering segment*, cluster heads assign timeslots to the member nodes using time division multiple access scheme. Member nodes of a cluster transmit data, receive acknowledgements from the cluster heads, and calculate their movements inside and outside of the cluster at their allocated timeslot. Thus, no extra timeslot is required to calculate nodes mobility. We also presuppose that every node are consistent in requisites of mobility and so, while a node moves out of a cluster there is a high probability of a further node entering into that cluster. Though, if a node moves into a new cluster and sends J-R message to cluster heads, the cluster heads does not assign the node a timeslot until any timeslot becomes free for moving a node out of this cluster.

Unwavering segment Pseudo -code

```
struct node
{ char[10] clusterarea;
int numberofclusters;
int clusterid;
struct node *next;
}*p;
void Unwavering segment (int TDMA,float mobilitycalculation, char*
newCHselection,int datapacket)
int noofnodes;
void main(){ int i; char[10] k; Unwavering segment (i,&k);}
Unwavering segment (int TDMA,float mobilitycalculation, char*
newCHselection,int datapacket)
{ int r; for (int f=1;i<=r;i++)
{ for(int j=1;j<=k;j++)
{ if(node[k][j] = senseevent){ node[k][j] sends data to CH calculates
dataenergyConsumption[k][j];
calculate recieveenergy;
 CH sends acknowledgement;
}else{ node[k][j] sends special packet calculate
specialenergyconsumpotion[k][j];
  calculate recieve energy CH[k];
  CH sends acknowledgement node[k][j]; }
if node[k][j] moves inside cluster k {
++countmoveinsidecluster[node[k][j]]
} if CH not recieve data from [k][j] {
delete node[k][j]; notify BS about [k][j];
} else if node[k][j] not recieve acknowledgement from CH
{ broadcast joint request
} if BS recieve new node and moved node for node[k][j]
{ mark node[k][j] as moved
 ++countmoveoutsidecluster[node[k][j]]
} else if CH recieve new node or move node
{ mark node[k][j] as failed;
} }
CH[k]aggregates data and sends to BS
CH energy consumption[k] for aggregating, sending to BS;
}
```

at any time the node N sense the subscribed events at its allocated timeslot, the node N sends data packet to cluster heads In case of no such sensed event of attention, the node N sends a tiny sized particular packet to notify cluster heads that it is unmoving alive or within the communication range of cluster heads. After receiving the data or individual packet cluster heads replies to N with an acknowledgment packet. If a cluster heads does not receive any data or special packet from N at its allocated timeslot the cluster heads assumes that the node N either has moved out of the cluster or failed.

Then cluster heads deletes the node N from its members list and also the timeslot allocated to N. cluster heads in addition notifies Base station the identification details of N. On the other hand, whenever x does not receive any acknowledgment packet from cluster head,N assumes that it is no longer attached to its cluster head due to mobility. Then N broadcasts a J-R packet and the cluster head that are within the communication range of N and in addition have free timeslot replies N with an A-J packet. Then N registers to the cluster of the Cluster Heads from which N receives the A-J packet with the maximum signal strength etc.

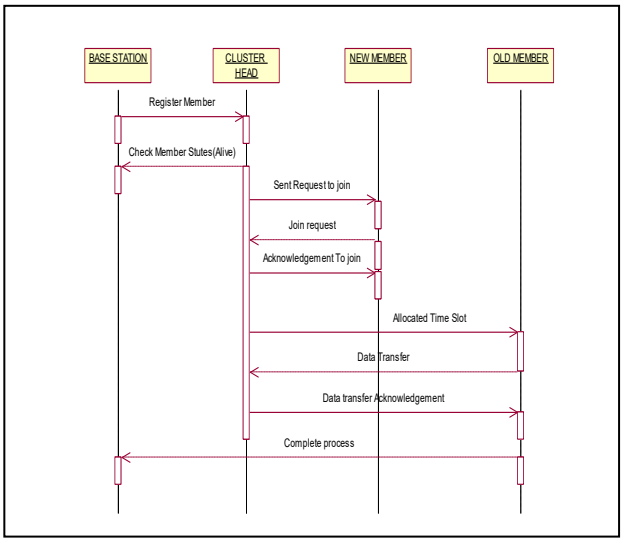


Figure: 2

IV. RESULTS SIMULATION

Simulation results show that the energy consumptions of the LEACH and DSC protocols are much more than that of the proposed ALWP protocol over a number of rounds Fig 3 demonstrates the network lifetime in terms of the remaining energy of the network over a number of rounds. Over research We discover that the energy debauchery in the LEACH and DSC protocol over rounds is much more than ALWP protocol and hence, the LEACH and DSC protocol have less network lifetime than that ALWP protocol.

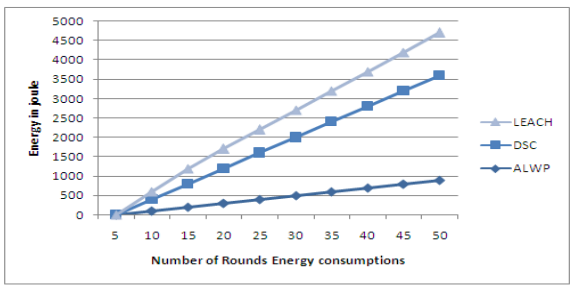


Figure : 3

V. CONCLUSION AND FUTURE WORKS

In this research we propose a new energy efficient routing algorithm which is related with distance factors and energy. Characteristic routing techniques have major purpose of boost the lifetime of the sensor network while not com promising data delivery. For the future work .we will propose sensor mediators going to use to sense, monitor and verification the concerned data. All of the cleverness mediators are going to be man-aged from a supervision center which is associated to descend by the satellite simulation.

ACKNOWLEDGMENT

We would like to express our gratitude to all those who gave us the possibility to complete this paper. We want to thank the Department of Engineering of the JJT University for giving me permission to commence this paper in the first instance, to do the necessary research work and to use departmental data. We are deeply indebted to our supervisor Prof. Dr. Lajish V.L from the JJT University whose help, stimulating suggestions and encouragement helped me in all the time of our research work for our Phd. and writing of this paper.

REFERENCE

[1]. List Akhan Akbulut, Cihangir, A. Halim Zaim, Güray Yılmaz," Energy and Distance Factor Based Routing Protocol for Wireless Sensor Networks Using Mobile Agents"- 2011 IEEE.

[2]. T.P. Lambrou, C.G. Panayiotou "A Survey on Rout-ing Techniques supporting Mobility in Sensor Net-works", Fifth International Conference on Mobile Ad-hoc and Sensor Networks 2009.

[3]. Lutful Karim and Nidal Nasser, "Energy Efficient and Fault Tolerant Routing Protocol for Mobile Sensor Network" IEEE ICC 2011.

[4]. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E.Cayirci. "Wireless sensor networks: a survey". Computer Networks, March 2002.

[5]. D. Braginsky, D. Estrin, "Rumor Routing Algorithm for Sensor Networks," Proc. 1st Wksp. Sensor Net-works and Apps., Atlanta, GA, sOct. 2002.

[6]. F. Bajaber and I. Awan, "Dynamic/Static Clustering Protocol for Wireless Sensor Network," Computer Modeling and Simulation, 2008. EMS '08. Second UKSIM European Symposium on, pp. 524-529, 2008.

[7]. F. Ye et al., "A Scalable Solution to Minimum Cost Forwarding in Large Sensor Networks,"

Proc. 10th Int'l. Conf. Comp. Commun. And Networks, 2001, pp. 304–09.

[8]. Suchetana Chakraborty, Sandip Chakraborty, Sukumar Nandi, Sushanta Karmakar," A Reliable and Total Order Tree Based Broadcast in Wireless Sensor Network" International Conference on Computer & Communication Technology (ICCCT)-2011.

[9]. Do-Seong Kim, Yeong-Jee Chung, "Self-Organization Routing Protocol Supporting Mobile Nodes for Wireless Sensor Network", Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06), 2006.

[10]. S. A. B. Awwad, C. K. Ng, N. K. Noordin and M. F. A. Rasid, "Cluster based routing protocol for mobile nodes in wireless sensor network," in Collaborative Technologies and Systems, 2009. CTS '09. International Symposium on, 2009, pp. 233-241.

[11]. J. Luo and J.-P. Hubaux, "Joint mobility and routing for lifetime elongation in wireless sensor networks," Proceeding of 24th Annual JointConference of the IEEE Computer and Communications Societies, 2005,pp. 1735-1746.

[12]. S.S. Kulkarni and M. Arumugam, "TDMA service for sensor networks,Proceedings. 24th International Conference on Distributed Computing System Workshops, 2004, pp. 604-609. "2.4 GHz IEEE 802.15.4 / ZigBee-Ready RF Transceiver (Rev. B)," Texas Instruments, 2000

[13]. W. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-Efficient Communication Protocol for Wire-less Microsensor Networks," Proc. 33rd Hawaii Int'l. Conf. Sys. Sci., Jan. 2000.

[14]. S. Lindsey, C. Raghavendra, "PEGASIS: Power-Efficient Gathering in Sensor Information Systems," IEEE Aerospace Conf. Proc., 2002, vol. 3, 9–16, pp. 1125–30.

[15]. A. Manjeshwar, D. P. Agarwal, "TEEN: a Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks," 1st Int'l. Wksp. On Parallel and Distrib. Comp. Issues in Wireless Networks and Mobile Comp., April 2001

◈◈◈

Analysis of Social Networking Sites Using K- Mean Clustering Algorithm

D. S. Rajput[1], R. S. Thakur[2] G. S. Thakur [3] & Neeraj Sahu [4]
[1,2&3]Department of Computer Applications, MANIT, Bhopal (MP), India
[4]Singhania University Rajasthan

***Abstract* -**Clustering is one of the very important technique used for classification of large dataset and widely applied to many applications including analysis of social networking sites, aircraft accidental, company performance etc. In recent days, Communication, advertising through social networking sites are most popular and interactive strategy among the users. This research attempts to find the large scale measurement study and analysis, effectiveness of communication strategy, analyzing the information about the usage, people's interest in social network sites in promoting and advertising their brand in social networking sites. The significance of the proposed work is determined with the help of various surveys, and from people who use these sites. Further a more specific pre-processing method is applied to clean data and perform the clustering method to generate patterns that will be work as heuristics for designing more effective social networking sites.

***Keywords* -** *Knowledge discovery, K-mean clustering, pre-processing, nearest neighbour searching, social networking sites.*

I. INTRODUCTION

We begin with a brief overview of social networks sites. Internet is primarily a source of communication, information and entertainment. In recent years; the use of social networking sites has been increasing. The use of these sites, interaction between the people is becoming easy. It is used by school colleges and IT professionals etc. It is important to understand why people use these websites; some people use them for business purposes, find new deals, legal and criminal investigations etc. Few social networking sites such as Facebook(2004), Twitter (2006), Myspace (2003), Orkut (2004), Friendster (2002), hi5 (2003), Google+(2011) etc, where people are connected with others directly.

Cluster analysis is a popular statistical tool for finding groups of respondents, objects, or cases that are similar to one another but different from those in other groups [1,2,11]. Analysis of social networking sites is closely dependent on clustering algorithms. There are many existing clustering algorithms such as K-Means, Fuzzy C-Means (FCM), CLERA, PAM, CLERANS etc [1,2,3,8,10,15] have their own pros and cons. K-Means is very fast but its center value is dependent on the initial assumptions K-means clustering (k-means), simply speaking, is an algorithm to classify or to group objects based on attributes or features into k number of group [2,5,13,14,16]. The grouping is done by minimizing the distances between data and the corresponding cluster centroid. In the application of means, we need to decide the value of k before starting the program, it should be noticed that different value of k will cause different levels of accuracy of the grouping.

II. LITERATURE REVIEW

Tapas Kanungo at el. in 2002 proposed An Efficient k-Means Clustering Algorithm: Analysis and Implementation [5]. This paper presents a simple and efficient implementation of Lloyd's k-means clustering algorithm, which is called filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure.

Gengxin Chen at el. in 2003 proposed Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data [6]. This paper given embryonic stems cell gene expression data, and applied several indices to evaluate the performance of clustering algorithms. This study may provide a guideline on how to select suitable clustering algorithms and it may help raise relevant issues in the extraction of meaningful biological information from microarray expression data.

Xu Yang at el. in 2010 proposed K-Means Based Clustering on Mobile Usage for Social Network Analysis Purpose [4]. This work studied data mining for social network analysis purpose, which aims at finding people's social network patterns by analyzing the information about their mobile phone usage.

Xiaoyan Li at el. in 2011 proposed Hybrid Retention Strategy Formulation in Telecom Based on k-means Clustering Analysis [7]. This paper proposed an idea of

formulating hybrid strategy to retain valuable customers using clustering technology.

Objectives of the study:-

I. To analyze easily the various conditions responsible for the various social networking site used by the youth.

II. To analyze the great help at which networking site is mostly used.

III. To analyze the effective communication strategy through social networking sites.

IV. To study the effectiveness of brand communication through social networking sites from its users and communicators.

V. To find the impact of interaction through these communication among Indian.

III. PROPOSED METHODOLOGY

In this proposed work Analysis of social networking sites is totally dependent on clustering algorithms. The existing clustering algorithm K-mean is very fast algorithm. This algorithm is used to classify features into k number of group or to group objects based on attributes. The grouping is done by minimizing the similarity between data and the corresponding cluster centroid.

This section presents data collection, data preprocessing and clustering methodology to generate clusters. There are many friends and relatives who are on facebook, Skype, Google+, Orkut etc. First open the friend's page and save the HTML document. This procedure repeat for user's profile. Finally we have documents available for classification in unstructured format.

The fig.1 shows proposed framework which consists three modules.

a) Text pre-processing module

b) Clustering algorithm module.

c) Cluster extraction and Specific result module. This framework will receive input from unstructured HTML data. The first module will perform pre-processing and extract document set D, and second module will perform K- mean clustering technique to generate clusters. Finally the last module provides results analysis.

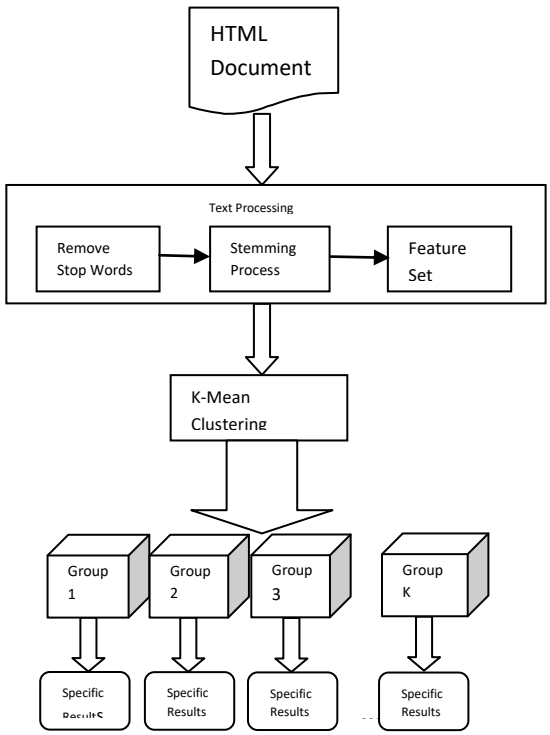


Figure 1: Framework for HTML Document

For the K-mean algorithm we have to decide value of K when beginning algorithm starts, it is noticed that different value of k will cause different levels of accuracy of the grouping[2,3,5,13,17]. The basic steps of K-mean clustering algorithm are:

Input: Number of cluster K.

Preprocessed Dataset.

1. Start
2. {
3. Take k samples from total number of N randomly as the centroid of each cluster.
4. Now Calculate the D of the remaining N-k sample to each centroid, and assign them to the cluster with the nearest centroid.
5. }
6. After each assignment, again calculate the centroid of the attainment cluster.
7. Now go to step 2 until find no new assignment.
8. Stop

IV. EXPERIMENTAL RESULTS

The proposed approach algorithm is applied in a social networking site dataset to generate clusters. To

explore the behavior of database we have choose the fields (Age Group, Time-spend, sex, occupation, Type of social site etc.). Aiming at the social site uses habits of Indian people; we have collected the data from the survey (like Family, Friends circle, college students etc.). All the experiments for finding the clustering results are performed on Pentium 2.6 GHz Processor, 2 GB RAM, Microsoft Windows 7.

The input means plots found in the Cluster node Results display the input means for the variables that were used in the clustering analysis over all of the clusters. The input means are normalized using a scale transformation function:

$$Y=(X-\min(X))/ (\max(x))-(\min(x))$$

For example, assume 5 input variables $Y_i= Y_1,...,Y_5$ and 3 clusters C_1, C_2, and C_3. Let the input mean for variable Y_i in cluster C_j be represented by M_{ij}. Then the normalized mean, or input mean, SMij becomes

$$SM_{ij}=(M_{ij}-\min(M_{i1},M_{i2},M_{i3})/((\max(M_{i1},M_{i2},M_{i3}))-(\min(M_{i1},M_{i2},M_{i3})))$$

The initial seeds must be complete cases, that is, training cases that have no missing values, and are required to be separated by a Euclidean distance that is of at least the value specified for the Minimum distance between cluster seeds (radius). By default, the initial seeds are chosen to be as far apart as possible; that is, seed replacement is set to full.

Using the data of social networking, the reproducibilities were as follows, for Initial seeds:

Maxclusters=40 Maxiter=1 and it is shown by Table.1

Cluster	1	2	3	4	5	6	7	8
1	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
6	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
8	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 1: Initial seeds

After first iteration we find the cluster Summary that has 8 clusters and find their frequency and distance between cluster centroids, which is shown by Table.2.

The Statistics tab of the Clustering Results Browser displays a table of clustering statistics Table.3 produced by the node's underlying vector quantization procedure. The entire training data set is used to calculate the following statistics for each cluster.

The root-mean-square standard deviation (RMS STD) measures the homogeneity of the cluster formed at any given step. It essentially measures the compactness or homogeneity of a cluster. Clusters in which consumers are very close to the centroid are compact clusters. The smaller the RMS STD, the more homogeneous or compact is the cluster formed at a given step. A large value of RMS STD suggests that the cluster obtained at a given step is not homogeneous, and is probably formed by merging of two very heterogeneous clusters. the first cluster would have a low RMSSTD whereas the second cluster would have a high RMSSTD. Notice that the cluster with a low RMSTD is relatively more homogeneous than the cluster with a high RMSTD. In general a cluster solution with a low RMSTD is preferred as it implies that the resulting clusters are homogeneous.

Cluster	Frequency	Root Mean Square Std Deviation	Maximum Distance from seed to Observation	Radius Exceeded	Nearest cluster	Distance between cluster centroids (mean)
1	33	0	0	-	8	1.4142
2	11	0	0	-	8	1.4142
3	15	0	0	-	8	1.4142
4	9	0	0	-	8	1.4142
5	10	0	0	-	8	1.4142
6	8	0	0	-	8	1.4142
7	6	0	0	-	8	1.4142
8	7	0	0	-	7	1.4142

Table 2: Cluster Summary after First Iteration

Variable	Total Standard Deviation	Within STD	R-Square	RSQ/(1-RSQ)
1	0.47380	0	1.000000	.
2	0.25764	0	1.000000	.
3	0.31587	0	1.000000	.
4	0.27393	0	1.000000	.
5	0.23982	0	1.000000	.
6	0.36037	0	1.000000	.
7	0.28894	0	1.000000	.
8	0.30288	0	1.000000	.
OVER ALL	0.32177	0	1.000000	.

Table 3: Statistics for Variables

*R-Square for predicting the variable from the cluster

*RSQ/(1-RSQ), which is the ratio of between-cluster variance to within cluster variance

Approximate Expected Over-All R-Squared = 0.52901

WARNING: The two values above are invalid for correlated variables.

In these results we find the cluster Mean. After final seed iteration Table.4 we find the final cluster Summary that has 6 clusters and find their frequency and distance between cluster centroids.

Cluster	Frequency	RMS Std Deviation	Maximum Distance from seed to Observation	Radius Exceeded	Nearest cluster	Distance between cluster centroids
1	46	0.2385	1.1375	-	2	1.2469
2	11	0	0	-	1	1.2469
3	15	0	0	-	1	1.2469
4	9	0	0	-	1	1.2469
5	10	0	0	-	1	1.2469
6	8	0	0	-	1	1.2469

Table 4: Cluster Summary after final Seed

A. *Distance between clusters:-*

It is shown in fig.2. The graph axes are determined from multidimensional scaling analysis, using a matrix of distances between cluster means as input. Therefore, it may appear that clusters overlap, but in fact, each case is assigned to only one cluster. The distance among the clusters is based on the criteria that are specified to construct the clusters. For illustrating clean the distance between clusters also shown by Table 5.

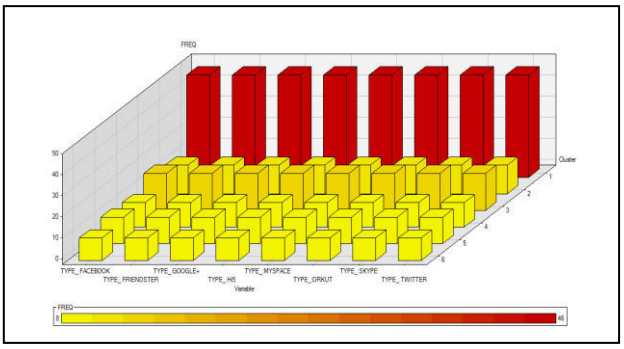


Figure 2: Distance between Clusters

Cluster	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
1	0	1.2469	1.2469	1.2469	1.2469	1.2469
2	1.2469	0	1.2469	1.4142	1.4142	1.4142
3	1.2469	1.4142	0	1.4142	1.4142	1.4142
4	1.2469	1.4142	1.4142	0	1.4142	1.4142
5	1.2469	1.4142	1.4142	1.4142	0	1.4142
6	1.2469	1.4142	1.4142	1.4142	1.4142	0

Table5: Distance between Clusters

B. *Cluster Tree:-*

Cluster tree in fig.3 displays a decision tree (path) for selected cluster. The decision tree is based on the sample of the training data set that was configured in the Clustering node configuration interface, specifically in the Preliminary Training and Profiles subtab of the Data tab. The cluster variable is used as the target variable, and the tree enables us to identify influential inputs. The fig.3 shows the all node portion of the tree for all clusters.

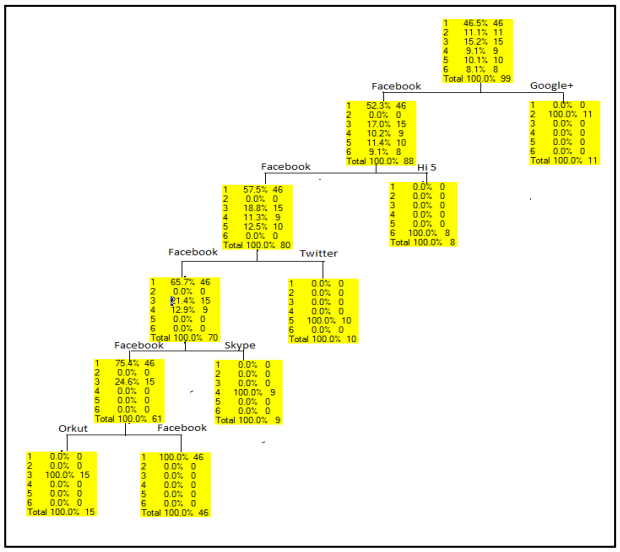


Figure 3: Clusters Tree

V. CONCLUSIONS

This paper analyzed social networking sites data using K-mean classification algorithm. The experimental results of text document classification on social networking sites dataset shows that 46% user preferred Facebook, 15% user preferred Orkut, 11% user preferred Google+, 10% user preferred Twitter, 9% user preferred Skype, 8% user preferred Hi5. This analysis concludes that the most common used website is the facebook.

1. By this analysis we can easily understand the various conditions responsible for the various social networking sites used by the youth.
2. The analysis is a great help at which networking site is mostly used.
3. This analysis also shows that this method works efficiently, for large text data.

ACKNOWLEDGMENT

This work is supported by research grant from MPCST, Bhopal M.P., India, Endt.No. 2427/ CST/R&D/2011 dated 22/09/2011.

REFERENCES

[1] Bryan Orme and Rich Johnson, Sawtooth Software, "Improving K-Means Cluster Analysis: Ensemble Analysis Instead of Highest Reproducibility Replicates" in 2008.

[2] Han I and Kamber M, "Data Mining concepts and Techniques,"Morgar Kaufmann Publishers,2000.

[3] D. Boley. Principal direction divisive partitioning. Data Mining and Knowledge Discovery, 2(4):325–344, 1998.

[4] Xu Yang, Yapeng Wang, Dan Wu, Athen Ma, "K-Means Based Clustering on Mobile Usage for Social Network Analysis Purpose" IEEE in 2010, pp 223-228.

[5] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002 ,pp 881-892.

[6] Gengxin Chen, Saied A. Jaradat, Nila Banerjee, Tetsuya S. Tanaka, "Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data" in 2003 pp 1-33.

[7] Xiaoyan Li, Yang Huang, Shujuan Li and Yishi Zhang, "Hybrid Retention Strategy Formulation in Telecom Based on k-means Clustering Analysis" IEEE in 2011, pp 978-981.

[8] Soon, M. C. , John, D. H., and Yanjun, L., "Text document clustering based on frequent word meaning sequences," Data & Knowledge Engineering, vol. 64, pp. 381-404, 2008.

[9] Chun-Ling Chen , Frank S.C. Tseng , Tyne Liang "An integration of WordNet and fuzzy association rule mining for multi-label document clustering" Data & Knowledge Engineering 69 (2010) pp 1208–1226 .

[10] Julie Beth Lovins, "Development of a Stemming Algorithm" Mechanical Translation and Computational Linguistics, vol.11, nos.1 and 2, March and June 1968.pp. 22-31.

[11] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, John Wiley & Sons, Inc, 2005. International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010.

[12] D.R. Recupero, "A new unsupervised method for document clustering by using WordNet lexical and conceptual relations, Information Retrieval" 10 (6) (2007) pp 563–579.

[13] Bryan Orme & Rich Johnson, "Improving K-Means Cluster Analysis: Ensemble Analysis Instead of Highest Reproducibility Replicates"in 2008.

[14] Andreas Hotho, Andreas N¨urnberger, Gerhard Paa, "A Brief Survey of Text Mining", in 2005.

[15] R.Vishnu Priya, A.Vadivel, R.S.Thakur, "Frequent Pattern Mining using Modified CP-Tree for Knowledge Discovery", Advanced Data Mining and Applications, LNCS-2010 volume 6440, pp. 254–261, © Springer-Verlag Berlin Heidelberg 2010

[16] R.S. Thakur, R.C. Jain, K.R. Pardasani, "Method of Conjugate Gradient: A Numerical Method for Mining Knowledge from Technical Data", in Research Hunt An International Multi Disciplinary Journal, Bhopal Vol. 1(1):182-189, 2006

[17] D.S. Rajput, R.S. Thakur, G.S. Thakur "Rule Generation from Textual Data by using Graph Based Approach" International journal of computer application, (IJCA) 0975 – 8887, New york USA, ISBN: 978-93-80865-11-8, Volume 31– No.9, October 2011.

◈◈◈

Clustering Based Classification and Analysis of Data

Neeraj Sahu[1] , D. S. Rajput[2], R. S. Thakur[3], G. S. Thakur[4]
[1] Singhania University Rajasthan
[2,3&4] Department of Computer Applications, MANIT, Bhopal (MP), India

***Abstract* -**This paper presents Clustering Based Document classification and analysis of data. The proposed Clustering Based classification and analysis of data approach is based on Unsupervised and Supervised Document Classification. In this paper Unsupervised Document and Supervised Document Classification are used. In this approach Document collection, Text Pre-processing, Feature Selection, Indexing, Clustering Process and Results Analysis steps are used. Twenty News group data sets [20] are used in the Experiments. For experimental results analysis evaluated using the Analytical SAS 9.0 Software is used. The Experimental Results show the proposed approach out performs.

***Keywords* -** *Document clustering, Unsupervised learning, Supervised learning, Text Pre processing*.

I. INTRODUCTION

Document Clustering is an important issue in text mining. Clustering has been widely applicable in different areas of science, technology, social science, biology, economics, medicine and stock market. Clustering problem appears in other different field like pattern recognition, statistical data analysis, bio-informatics, etc.There exist Clustering based classification in the literature. Clustering based classification are mainly divided into two categories shown in Fig.1:

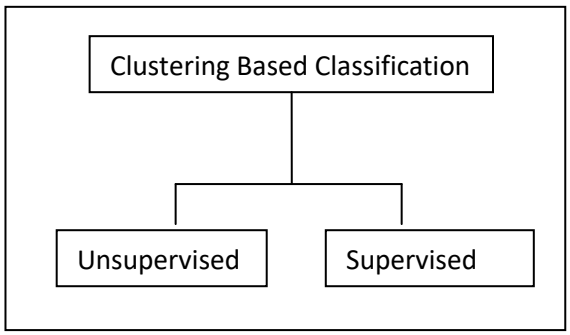


Fig. 1 : Tree structure with Clustering based classification

In recent years lot of research work has been done on Document Clustering. Some contributions are as follows:

In 2002 Beil et al.[1] worked to improve the cluster accuracy using frequent item based technique and find overlapping clusters and meaning full cluster label.

In 2010 Chun-Ling Chen,Frank S.C. Tseng,Tyne Liang[4],Fuzzy-based Multi-label Document Clustering (FMDC) algorithm concentrated on clustering accuracy and used frequent item based clustering concept and find overlapping cluster, semantic discovery and meaningful cluster label.The above mentioned work suffers from lack of efficiency and accuracy. The high complexity and low accuracy are still issues and challenges in the clustering. This motivates the study of Document Clustering.

The paper is organized as follows. Section-I described the introduction and review of literatures. In Section-II, the Unsupervised and Supervised Document Classification are described. In Section-III, Methodology of document clustering is described. In Section-IV, Experimental results are described. In Section-V, Evaluation measurement is described. Finally, we concluded and proposed some future directions in Conclusion Section.

II. UNSUPERVISED AND SUPERVISED CLASSIFICATION

Document clustering has been used to enhance information retrieval. This is based on the clustering hypothesis, which states that documents having similar contents are also relevant to the same query [1]. A fixed collection of text is clustered into groups or clusters that have similar contents. The similarity between documents is usually measured with the associative coefficients from the vector space model, e.g., the cosine coefficient. Hierarchical clustering algorithms have been primarily been used in document clustering. The single link method has mostly been used, as it is

computationally feasible, but the complete link method seems to be the most effective but is very computationally demanding [2].

Other methods than document vector similarity have been used for clustering. Neural models have been implemented for unsupervised document clustering [3].

The long computation time has always been the problem when using document clustering on-line. More recently fast algorithms for clustering have been introduced to use for browsing through collection when the user has little information about the collection and wants to brows for topics [4]. Suffix Tree Clustering is new clustering method, which creates clusters based on phrases shared between documents, works fast and intended for Web document clustering [5]. Different projections techniques, LSI and truncation, have been investigate to speed up the distance calculations of clustering [6]. An interesting application of clustering is topic clustering, i.e. clustering documents returned from a specific query, using -means clustering [7]. Effectiveness of five hierarchical clustering algorithms have been examined: single link, complete link, group average, Ward's method, and weighted average [8]. Single link is the only that badly compared to the others, but the results are very much dependent on the data set.

Supervised Document Classification:- Pattern recognition and machine learning has also been applied to document classification. As before the term frequency is used as feature. A number of classifiers have been used to classify documents. An example of these classifiers are neural networks [9,10], support vector machines [11], genetic programming [12], Kohonen type self-organizing maps [13], hierarchically organized neural network built up from a number of independent self-organizing maps [14], fuzzy -means [15], hierarchical Bayesian clustering [16], Bayesian network classifier [17], and naive Bayes classifier [18].

Some of these classifiers can be used with unsupervised learning, i.e., unlabeled documents, but the accuracy of a classifier can be enhanced by using a small set of labeled documents [18]. The aim is to use a classifier which need small amount of manually classified documents to be generalized.

The use of semi-supervised machine learning has emerged recently [18,15]. The learning scheme lies somewhere between supervised and unsupervised, where the class information are learned from the labeled data and the structure of the data from the unlabeled data.

The performance of four document classification methods have been measured: the naive Bayes classifier, the nearest classifier, decision trees and a subspace method [19]. The naive Bayes classifier and the subspace method outperform the others.

1. Supervised: In Supervised classification method, a set of predefines classes are given.
2. Unsupervised: In Unsupervised classification methods, a set of predefine classes are not given. This is also known clustering.

III. METHODOLOGY

A. Document Collection

In this phase we collect relevant documents like e-mail, news, web pages etc. from various heterogeneous sources. These text documents are stored in a variety of formats depending on the nature of the data. The datasets are downloaded from UCI KDD Archive [20]. This is an online repository of large datasets and has wide variety of data types.

B. Text Pre-processing

Text pre-processing means transform documents into a suitable representation for the clustering task. The text documents have different stop words, punctuation marks, special character and digits and other characters. After removing stop words, word stemming is performed .Word stemming is the process of suffix removal to general word stems. A stem is a natural group of words with similar meaning. In text-pre-processing we performed the following task:

a) Removal of HTML tags and special character

b) Removal stop words

c) Word stemming

C. Dimension reduction

High dimension is the greatest challenge of document clustering, so dimension reduction became major issue for clustering. This module performs two functions- indexing and feature selection. In indexing method we assign the value to the terms in the documents. After indexing, feature selection method is applied. Feature selection is the process of removing indiscriminate terms from the documents to improve the document clustering accuracy and reduce the computational complexity.

D. Word Stemming

Morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of Information Retrieval applications. For this reason we have been developed, which attempt to reduce a word to its stem or root form. The key terms of a document are represented by stems

rather than by the original words. This not only means that different variants of a term can be conflated to a single representative form – it also reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time.

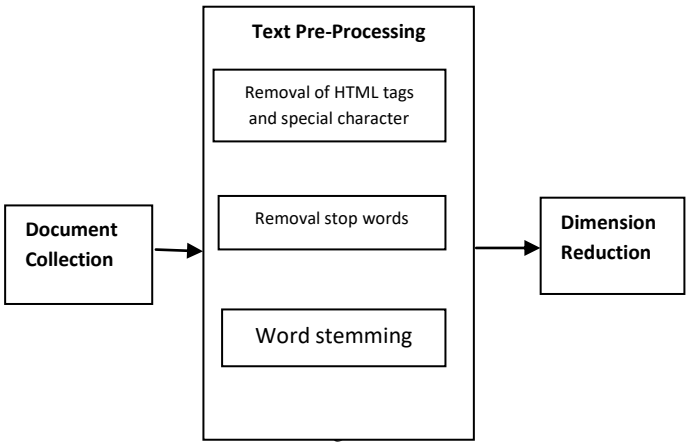


Figure 2. Steps of Methodology

E. Stop Words

Stop Word list is probably the most widely used stop word list. It covers a wide number of stop words without getting too aggressive and including too many words which a user might search upon. This wordlist contains 429 words it is given below:

about,above,across,after,again,against,all,almost,alone,along,already,also,although,always,among,an,and,another,any,anybody,anyone,anything,anywhere,are,area,areas,around,as,ask,asked,asking,asks,at,away,back,backed,backing,backs,be,became,because,become,becomes,been,before,began,behind,being,beings,best,better,betweenbig,both,but,by,came,can,cannot,case,cases,certain,certainly,clear,clearly,comecould,did,differ,different,differently,do,does,done,down,down,downed,downing,downs,during,each,early,either,end,ended,ending,ends,enough,even,evenly,everevery,everybody,everyone,everything,everywhere.

face,faces,fact,facts,far,felt,few,find,finds,first,for,four,from,full,fully,further,furthered,furthering,furthers,gave,general,generally,get,gets,give,given,gives,go,going,good,goods,got,great,greater,greatest,group,grouped,grouping,groups,had,has,have,having,he,her,here,herself,high,high,high,higher,highest,him,himself,his,how,however,if,important,in,interest,interested,interesting,interests,into,is,it,its,itself,just,keep,keeps,kind,knew,know,known,knows,large,largely,last,later,latest,least,lesslet,lets,like,likely,long,longer,longest,made,make,making,man,many,may,me,membermembers,men,might,more,most,mostly,mr,mrs,much,must,my,myself,necessary,need,needed,needing,needs,never,new,new,newer,newest,next,no,nobody,non,no,one,not,nothing,now,nowhere,number,numbers,of,off,often,old,older,oldest,on,once,one,only,open,opened,opening,opens,or,order,ordered,ordering,orders,other,others,our,out,over,part,parted,parting,parts,per,perhaps,place,places,point,pointedpointing,points,possible,present,presented,presenting,presents,problem,problems,put,puts,quite,rather,really,right,right,room,rooms,said,same,saw,say,says,second,seconds,see,seem,seemed,seeming,seems,sees,several,shall,she,should,showshowed,showing,shows,side,sides,since,small,smaller,smallest,so,some,somebody,someone,something,somewhere,state,states,still,still,such,sure,take,taken,thanthat,the,their,them,then,there,therefore,these,they,thing,things,think,thinks,this,those,though,thought,thoughts,three,through,thus,to,today,together,too,tooktoward,turn,turned,turning,turns,two,under,until,up,upon,us,use,used,uses,very,want,wanted,wanting,wants,was,way,ways,we,well,wells,went,were,what,when,where,whether,which,while,who,whole,whose,why,will,with,within,without,work,worked,working,works,would,year,years,yet,you,young,younger,youngest,your,yours.

F.HTML Tags:

<!doctype>, <a>, <abbr>, <acronym>, <address>, <applet>, <area>, <b>, <base>, <basefont>, <bdo>, <big>, <blockquote>, <body>,
, <button>, <caption>, , <cite>, <code>, <col>, <colgroup>, <dd>, <del>, <dfn>, <dir>, , <dl>, <dt>, , <fieldset>, <font>, <form>, <frame>, <frameset>, <h1>, <h2>, <h3>, <h4>, <h5>, <h6>, <head>, <hr>, <html>, <i>, <iframe>, <img>, <input>, <ins>, <isindex>, <kbd>, <label>, <l,egend>, <li>, <link>, <map>, <menu>, <meta>, <noframes>, <noscript>, <object>, <ol>, <optgroup>, <option>, <p>, <param>, <pre>, <q>, <s>, <samp>, <script>, <select>, <small>, , <strike>, , <style>, <sub>, <sup>, <table>, <tbody>, <td>, <textarea>, <tfoot>, <th>, <thead>, <title>, <tr>, <tt>, <u>, <ul>, <var><!-->

IV. EXPERIMENTAL RESULTS

In this paper the unstructured datasets are used. The datasets are downloaded from UCI KDD Archive [20]. This is an online repository of large datasets with wide variety of data types. This repository has twenty newsgroups dataset for text analysis. This data set consists of 20000 messages taken from Usenet newsgroup. The subset of twenty newsgroups is mini newsgroup. We have done our experiments on 20 newsgroup datasets. Each category contains 1000 documents, so there are 20000 documents for experiments. The five categories Computer Hardware,Computer Graphics,Medical,Sports and Automobile are used in first experiment.

We performed our experiments on five newsgroups- Computer graphics, Computer hardware, Automobile, Sports and Medical. In this research the 80% dataset are used as training dataset and 20% dataset are used as test dataset.

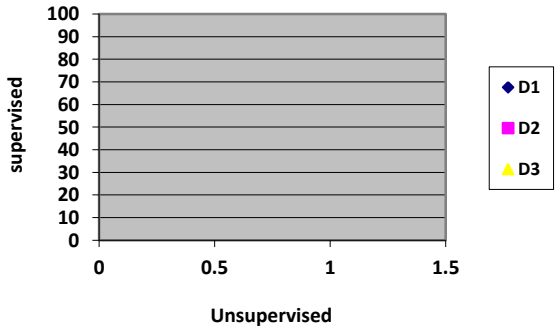


Fig.2 D1, D2 , D3 Data Set with Clustering based classification

ACKNOWLEDGMENT

This work is supported by research grant from MPCST, Bhopal M.P., India, Endt.No. 2427/CST /R&D/2011 dated 22/09/2011.

REFERENCES

[1] van Rijsbergen, C. J. Information retrieval. Butterworths, 1979.

[2] Willett, Peter. Recent Trends in Hierarchic Document Clustering: A Critical Review. Information Processing and Management Vol. 24, No 5, p. 577-597, 1988.

[3] MacLeod, K. An application specific neural model for document clustering. Proceedings of the Fourth Annual Parallel Processing Symposium, vol.1, p. 5-16, 1990.

[4] Douglass Cutting, David R. Karger, Jan O. Pedersen and John W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In Proceedings of ACM/SIGIR, p. 318-329, 1992.

[5] Zamir, O. and Etzioni, O. Web document clustering: a feasibility demonstration. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 46-54, 1998.

[6] Schutze, Hinrich and Silverstein, Craig Projections for efficient document clustering. SIGIR Forum (ACM Special Interest Group on Information Retrieval), p. 74-81, 1997.

[7] Sahami, Mehran; Yusufali, Salim and Baldonado, Michelle Q.W. Real-time full-text clustering of networked documents. Proceedings of the National Conference on Artificial Intelligence, p. 845, 1997.

[8] Burgin, R. The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. Journal of the American Society for Information Science, Vol.46 (8), p. 562-72, 1995.

[9] Li, Wei; Lee, Bob; Krausz, Franl and Sahin, Kenan. Text Classification by a Neural Network. Proceedings of the 1991 Summer Computer Simulation Conference.Twenty-Third Annual Summer Computer Simulation Conference,p. 313-318, 1991.

[10] Farkas, Jennifer. Generating Document Clusters Using Thesauri and Neural Networks. Canadian Conference on Electrical and Computer Engineering, Vol.2, p. 710-713, 1994.

[11] Joachims, Thorsten. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Machine Learning: ECML-98. 10th European Conference on Machine Learning, p. 137-42 Proceedings. 1998.

[12] Svingen, B. Using genetic programming for document classification. FLAIRS-98. Proceedings of the Eleventh International Florida Artificial Intelligence Research, p. 63-67, 1998.

[13] Hyotyniemi, H. Text document classification with self-organizing maps. STeP '96 - Genes, Nets and Symbols. Finnish Artificial Intelligence Conference, p. 64-72, 1996.

[14] Merkl, D. Text classification with self-organizing maps: Some lessons learned. Neurocomputing Vol.21 (1-3), p. 61-77, 1998.

[15] Benkhalifa, M., Bensaid, A. and Mouradi, A. Text categorization using the semi-supervised fuzzy c-means algorithm. 18th International Conference of the North American Fuzzy Information Processing Society - NAFIPS, p. 561-5, 1999.

[16] Iwayama, M. and Tokunaga, T. Hierarchical Bayesian Clustering for automatic text classification. IJCAI-95. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol.2, p. 1322-7, 1995.

[17] Lam, Wai and Low, Kon-Fan Automatic document classification based on probabilistic

reasoning: Model and performance analysis. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Vol.3, p. 2719-2723, 1997.

[18] Nigam, Kamal; Maccallum, Andrew Kachites; Thrun, Sebastian and Mitchell, Tom. Text Classification from Labeled and Unlabeled Documents using EM. To appear in the Machine Learning Journal 1999. Draft.

[19] Li, Y.H. and Jain, A.K. Classification of text documents. Computer Journal, 41 (8) , p. 537-46, 1998.

[20] www.kdd.ics.uci.edu

◈◈◈

Ownership Verification/Authentication Using Visual Cryptography Based Digital Watermarking

Jaishri Chourasia[1], Keyur Parmar[2], Sowmya Suryadevara[3] & Sonam Rathore[4]

[1,3 &4]MITS, Lakshmangarh, [2]SVNIT, Surat, India

Abstract - We have proposed ownership verification/authentication scheme using visual cryptography based digital watermarking. Due to the advantage of visual cryptography, watermark pattern would be present as a share and single share does not contain the complete information of the original watermark pattern. The scheme does not embed the watermark pattern directly into the host image, hence it would be difficult to detect or remove the watermark in an illegal way, and also the scheme maintains the quality of the host image. The experimental results show that watermarking scheme is transparent and robust after several attacks.

Keywords - *Digital watermarking, visual cryptography, secret sharing.*

I. INTRODUCTION

Watermarking is one of the most popular techniques in protecting copyrights of digital media. Watermarking is also important for several imaging applications: trusted camera, legal usage of images, news reporting, commercial image transaction. In each of these applications, it is important to verify that the image has not been manipulated and the image was originated by a specific user. The watermark embedded into the image can be extracted for the purpose of ownership verification and/or authentication. Due to the combination of the computer and communication technology, more and more digital documents are transmitted and exchanged over the internet. It has created an environment that digital information is easy to distribute, duplicate and modify. This has led to need for effective copyright protection technique [1].

The image watermarking method must satisfy the following requirements [2]:

1) Transparency: The embedded watermark pattern does not visually spoil the original image and should be perceptually invisible.

2) Robustness: The watermark pattern is hard to detect and remove in illegal way. It should be immune to various possible attacks.

As watermark is used for copyright protection it has become an attraction point to hack or break it in illegal way. Other watermarking techniques embedded watermark in the spatial domain and some in frequency domain [1], [2], [3], [4]. These schemes incorporate complex embedding technique and fails after attacks. Even if the watermark is invisible, watermark could be detected using any possible algorithm and can be misused. The proposed scheme does not embed the watermark pattern directly into the host image instead it uses the concept of visual cryptography [5], [6], [7] proposed by Naor and Shamir. The visual cryptography scheme divides the secret information such as printed text, image etc. into the parts called as shares. Any single share does not contain complete information about the secret information. To reconstruct the original secret information either all or subset of shares are required [8], [9]. We applied this concept for the watermarking scheme where the watermark pattern is divided into the shares. One share called as verification share is generated at the time of watermark embedding into the host image, called as marked image. Other called as master share is generated during watermark extraction process from the marked image [10]. The scheme ensures that the size and quality of the host image will remain after the embedding process.

The rest of the paper is organized as follows: Section II describes the (2, 2) visual cryptography scheme which is used for generation of shares of the watermark pattern. Section III describes the proposed watermarking scheme. Section IV shows experimental results. Finally, section V concludes the paper.

II. BRIEF DESCRIPTION OF (2, 2) VISUAL CRYPTOGRAPHY SCHEME

To encrypt the secret information using (2, 2) visual cryptography scheme, the secret information is dived into two shares such that each pixel in the original image is replaced with the non-overlapping block of two subpixels. Anyone who holds only one share will not able to reconstruct the secret information as single share does not contain complete secret information.

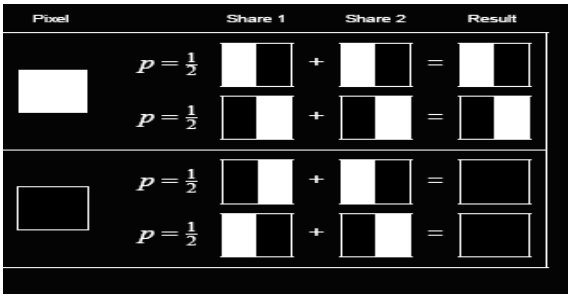


Fig. 1: (2, 2) Visual cryptography scheme

Figure 1 illustrates encoding scheme for (2, 2) visual cryptography which is to be applied on the every pixel of the secret information. If pixel p is white of the secret information then it is replaced with two identical blocks of subpixels. If the pixel p is black of the secret information then it is replaced with two complementary blocks of subpixels. To decrypt the secret information, each share is xeroxed onto the transparency. Superimposition of both transparencies will reveal the secret information.

III. THE PROPOSED WATERMARKING SCHEME

The proposed scheme is based on (2, 2) visual cryptography scheme described in the section 2. The owner of the host image should select a bi-level image as watermark pattern W of size P×Q. If W is not a bi-level image then it has to be converted into the bi-level image. In the embedding phase owner will generate the verification share (V) and in the extraction phase master share (S) using (2, 2) visual cryptography scheme. The owner requires registering the watermark pattern (P), secret key (K). While verifying rightful ownership verification share (V) and master share (S) will be used by notarial organization.

A. Embedding Scheme

In the embedding phase, the owner selects a random number 'K' as a secret key to embed the watermark pattern into the host image. The secret key could be same or different for different host images and it must be kept secretly. Let, the owner embed the watermark pattern (P) of size P×Q into the host image (I) of size X×Y. The embedding algorithm includes following steps:

Input : Secret key (K), host image (I) of size X×Y.

Output : Marked image (M) of size X×Y.

Step1. Select a random number K as the secret key of the host image (I).

Step2. Use 'K' as the seed to generate P×Q random numbers over the interval [1, h]; where h= X×Y. Let R_i is i^{th} random number.

Step3. Creation of binary matrix Z of size P×Q such that the elements of the matrix Z are the least significant bit of the R_i; i^{th} random number.

Step4. Assign the i^{th} pair (V_{i1}, V_{i2}) of the verification share (V) based on the information given in the table 1 using the pixel value of W and matrix Z.

Step5. Assemble all the pair values to construct the verification share (V) of size P×2Q.

B. Extraction and Verification/Authentication Scheme

If the owner wants to prove that somebody is imitating host image (I) as image 'O' the notarial organization can identify it by following steps:

Input: Marked image (M) of size X×Y, watermark pattern W of size P×Q, verification share (V) of size P×2Q.

Output: Master share (S) of size P×2Q, extracted watermark pattern (W').

Step1. Use secret key 'K' as a seed to generate P×Q random numbers over the interval [1, h]; where h=X×Y. Let R_i is i^{th} random number.

Step2. Creation of binary matrix Z' of size P×Q such that the elements of the matrix Z' are the least significant bit of the R_i; i^{th} random number.

Step3. Assign the i^{th} pair (S_{i1}, S_{i2}) of the master share (M) based on the information given in the table 2.

Step4. Assemble all the pair values to construct the master share (S) of size P×2Q.

Step5. To extract the watermark pattern compare verification share (V) and master share (S). If the i^{th} pair (V_{i1}, V_{i2}) of V is equal to the i^{th} pair (S_{i1}, S_{i2}) of S then assign the i^{th} element value of the extracted watermark pattern 1 else assign 0.

Step6.If through human visual system W' can be recognized as W, and then notarial organization shall adjudge that image 'O' is a copy of host image (I).

Table 1: The rules for generation of verification share

Color of the i^{th} pixel in watermark pattern W_i	i^{th} element in binary matrix Z_i	Pair of bits (V_{i1}, V_{i2}) to be assigned in verification share (V)
Black	1	(0, 1)
Black	0	(1, 0)
White	1	(1, 0)
White	0	(0,1)

Table 2: The rules for generation of master share

i^{th} element in binary matrix Z'_i	Pair of bits (S_{i1}, S_{i2}) to be assigned in master share (S)
1	(1, 0)
0	(0, 1)

IV. EXPERIMENTAL RESULTS

This section presents some experimental results concerning the proposed scheme. The proposed scheme is tested on the host image of the size 300×400 which is shown in the figure 2. The watermark pattern to be embedded is of size 15×46 shown in the figure 3. Figure 4 and figure 5 show the verification and master share. Figure 6 shows the extracted watermark pattern by the superimposition of the verification share and master share. All experiments are implemented in MATLAB Image Processing toolbox.

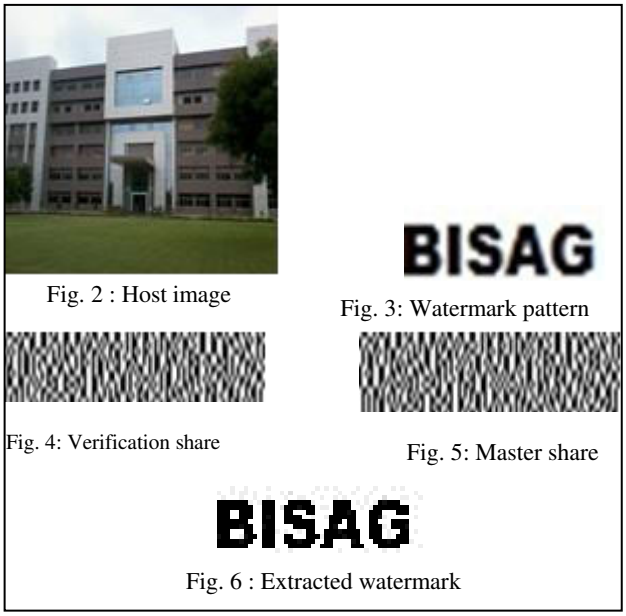


Fig. 2 : Host image

Fig. 3: Watermark pattern

Fig. 4: Verification share

Fig. 5: Master share

Fig. 6 : Extracted watermark

To test the robustness of the proposed scheme several watermarking attacks have been applied such as noise, sharpening, blurring, averaging filter and JPEG compression etc.

A. *Salt & Pepper Noise Attack*

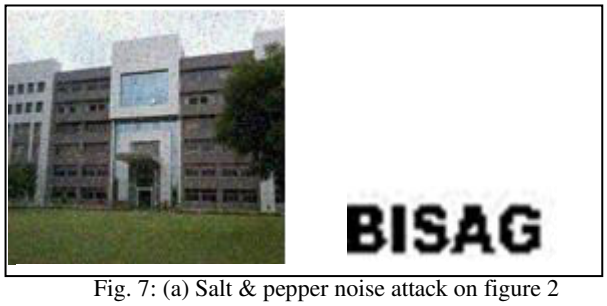


Fig. 7: (a) Salt & pepper noise attack on figure 2

B. *Sharpening Attack*

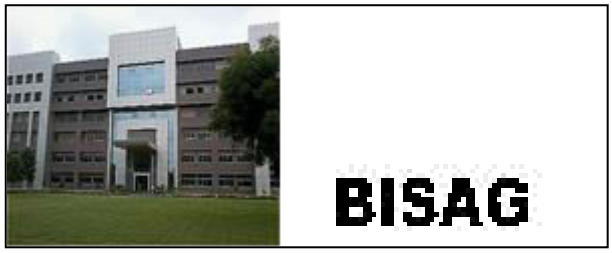


Fig.8 : (a) Sharpening attack on figure 2
(b) Extracted watermark

C. *Blurring Attack*

Fig. 9 : (a) Blurring attack on figure 2
(b) Extracted watermark

D. *Averaging Filter Attack*

Fig. 10: (a) Averaging filter attack on figure 2
(b) Extracted watermark

E. JPEG Compression Attack

Fig. 11: (a) JPEG compression attack on figure 2
(b) Extracted watermark

V. CONCLUSION

This scheme focuses on visual cryptography based digital watermarking. After performing several image processing experiments, this watermarking scheme seems to be robust against noise, sharpening, blurring and averaging distortions. The advantages of this scheme are:

(1) It does not require any complex transformations for embedding the watermark such as FFT, DCT etc.

(2) It maintains the quality of the host image after embedding the watermark.

(3) The watermark would be present as a share, which does not reveal complete information about the original watermark; also without the use of secret key it is impossible to extract the watermark pattern.

ACKNOWLEDGEMENT

This work was supported by Mr. Nisheeth Saxena, Mr. P. K. Bishnoi, MITS, Lakshmangarh (Rajasthan) and Mr. Abdul Jhummarwala, Dr. M.B. Potdar, BISAG, Gandhinagar (Gujrat), India.

REFERENCES

[1] Chunlin Song, Sud Sudirman and Madjid Merabti, “Robust Digital Image Watermarking using Region Adaptive Embedding Technique”, IEEE, pp. 378-382, 2010.

[2] Cox I. J., Kiliant J., Leighton T. and Shamoon T., ‘Secure Spread Spectrum Watermarking for Multimedia”, IEEE Transactions on Image Processing, Vol. 6, No. 12, pp. 1673-1687, 1997.

[3] Hsu C. and Wu J.,” DCT-Based Watermarking for Video”, IEEE Transactions on Consumer Electronics, Vol. 44, No.1, pp.206-215, 1998.

[4] Hsu C. T. and Wu J. L.,”Hidden Digital Watermarks in Images”, IEEE Transactions on Image Processng, Vol. 8, No. 1, Jan, pp. 58-68, 1999.

[5] A. Shamir,. “How to Share a secret”,Communications of the ACM, vol. 22, pp.612-613, 1996.

[6] M. Naor, and A. Shamir, “Visual Cryptography”, Advances in Cryptology – Eurocrypt’94 Proceeding, LNCS Vol. 950, Springer-Verlag, pp. 1-12, 1995..

[7] M. Naor and A. Shamir, “Visual Cryptography II:Improving the Contrast Via the Cover Base”,Cambridge Workshop on Protocols, 1996.

[8] Zhi Zhou,Gonalzo R. Arce and Giovanni Di Crescenzo, “Halftone Visual Cryptography”, IEEE Transactions On Image Processing, Vol. 15, NO. 8, pp. 2441-2453, 2006.

[9] Hao Luo, Faxin Yu. , “Data Hiding in Image Size Invariant Visual Cryptography”, IEEE pp. 273-276, 2008.

[10] Y-C Hou, P-M Chen, “An Asymmetric Watermarking Scheme based on Visual Cryptography”, In Proceedings of ICSP, Vol. 2, pp. 992 -995, 2000..

◈◈◈

A Novel Review on Routing Protocols in MANETs

Robinpreet Kaur & Mritunjay Kumar Rai

Department of Electronics and Engineering, Lovely Professional University, Phagwara, Punjab, India

Abstract - Mobile means moving and ad hoc means temporary without any fixed infrastructure so mobile ad hoc networks are a kind of temporary networks in which nodes are moving without any fixed infrastructure or centralized administration. MANETs are generating lots of interest due to their dynamic topology and decentralized administration. Due to the diverse applications which use MANETs for wireless roaming it is a current research issue. There are different aspects which are taken for research like routing, synchronization, power consumption, bandwidth considerations etc. This paper concentrates on routing techniques which is the most challenging issue due to the dynamic topology of ad hoc networks. There are different strategies proposed for efficient routing which claimed to provide improved performance. There are different routing protocols proposed for MANETs which makes it quite difficult to determine which protocol is suitable for different network conditions as proposed by their Quality of service offerings. This paper provides an overview of different routing protocols proposed in literature and also provides a comparison between them.

Keywords - *MANETs, routing protocol, reactive, proactive, hybrid, performance, dynamic topology.*

I. INTRODUCTION

In recent years MANET has gained popularity and lots of research is being done on different aspects of MANET. It is an infrastructure less network having no fixed base stations MANET is characterized by dynamic topology low bandwidth and low power consumption. All the nodes in the network are moving i.e. topology of the network is dynamic so the nodes can act both as host as well as router to route information unnecessary for its use. This kind of infrastructure-less network is very useful in situation in which ordinary wired networks is not feasible like battlefields, natural disasters etc. The nodes which are in the transmission range of each other communicate directly otherwise communication is done through intermediate nodes which are willing to forward packet hence these networks are also called as multi-hop networks

II. CHARACTERISTICS OF MANETs

Dynamic topology: Nodes are free to move arbitrarily in any direction thus the topology of the network change unpredictably.

Limited Bandwidth: the bandwidth available for wireless networks is generally low than that of wired networks. The throughput of these networks is generally low due various noises, fading effects.

Energy constrained operation: the nodes are portable devices and are dependent on batteries. This is the most important design consideration of the MANET

Security: wireless networks are more prone to threats than wired networks. The increased possibility of various security attacks like eavesdropping, denial of service should be handled carefully.

Performance of MANET depends on the routing protocol, battery consumption by the nodes. There are various Quality of service parameters which affect the performance like bandwidth delay, jitter, throughput etc. Due to dynamic topology routing is the major challenge in these networks because the bandwidth provided to the nodes at one point of time becomes unavailable if the nodes move from a particular position and go to other position. Moreover routing affects the performance of these networks. Therefore efficient routing protocol needs to be developed to meet all these challenges. routing protocol in MANET is classified into three categories on the basis of route discovery reactive also called as on demand routing protocol ,proactive also known as table driven protocol and Hybrid protocol. Further classification of routing protocols is done on the basis of network organisation as flat based, hierarchical based and location based. In flat based protocol all the nodes are equal i.e. they play the same role in the network. In hierarchical protocol different nodes play different roles i.e. in this different cluster heads are chosen among cluster members. In location based protocol nodes rely on the location information and use this information for communication.

III. ROUTING PROTOCOLS

Routing protocols define a set of rules which governs the journey of message packets from source to destination in a network. In MANET, there are different types of routing protocols each of them is applied

according to the network circumstances. Figure 1 shows the basic classification of the routing protocols in MANETs.

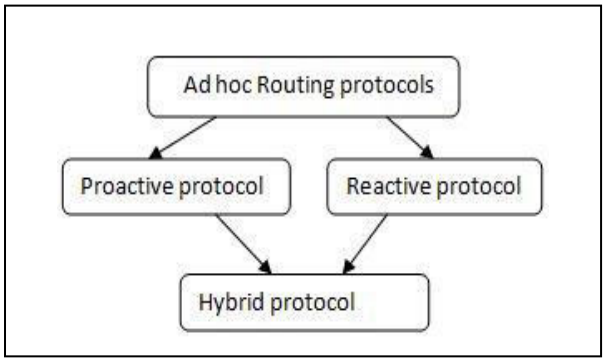


Fig. 1 : Classification of Routing protocols

i. Reactive Routing Protocols

Reactive routing protocol is also known as on demand routing protocol. In this protocol route is discovered whenever it is needed Nodes initiate 0route discovery on demand basis. Source node sees its route cache for the available route from source to destination if the route is not available then it initiates route discovery process. The on- demand routing protocols have two major components [1]:

Route discovery: In this phase source node initiates route discovery on demand basis. Source nodes consults its route cache for the available route from source to destination otherwise if the route is not present it initiates route discovery. The source node, in the packet, includes the destination address of the node as well address of the intermediate nodes to the destination.

Route maintenance: Due to dynamic topology of the network cases of the route failure between the nodes arises due to link breakage etc, so route maintenance is done. Reactive protocols have acknowledgement mechanism due to which route maintenance is possible

Reactive protocols add latency to the network due to the route discovery mechanism. Each intermediate node involved in the route discovery process adds latency. These protocols decrease the routing overhead but at the cost of increased latency in the network. Hence these protocols are suitable in the situations where low routing overhead is required.

There are various well known reactive routing protocols present in MANET for example DSR, AODV, TORA and LMR.

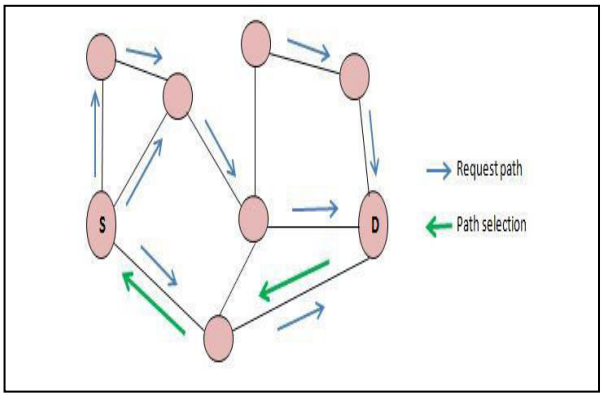


Fig. 2 : DSR protocol

In *Dynamic Source Routing (DSR)*, shown in Figure.2, the protocol is based on the link state algorithm in which source initiates route discovery on demand basis. The sender determines the route from source to destination and it includes the address of intermediate nodes to the route record in the packet. DSR was designed for multi hop networks for small Diameters.

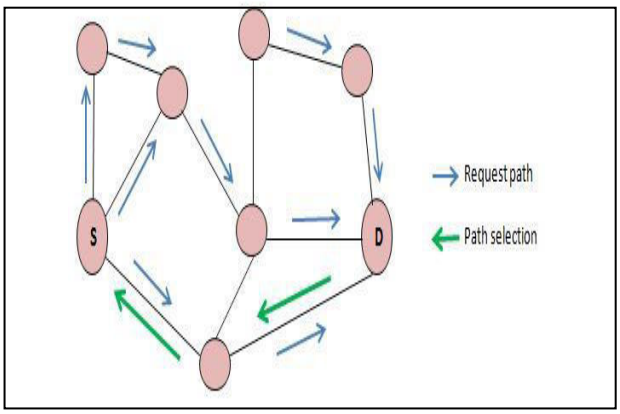


Fig. 3 : AODV protocol

It is a beaconless protocol in which no HELLO messages are exchanged between nodes to notify them of their neighbours in the network. *Ad hoc On Demand distance Vector (AODV)* is also a reactive routing protocol. In this protocol, instead of containing information about the complete network topology sender only includes the address of its neighbour in the packet. In this way overhead in this protocol is comparatively less than DSR. A basic AODV protocol is shown in Figure.3.*Temporally ordered routing algorithm (TORA), Light weight Mobile routing (LMR)* is also reactive protocol based on the link reversal algorithm. It also consists of two phases like DSR route establishment and route maintenance. In route establishment route is discovered by the use of query packets in the network, the route maintenance is done by sending failure query messages to detect route

failures in the network. There are various advantages as well as disadvantages of reactive protocols. As these are based on route discovery on demand bases so these include less overhead of control messages hence saving bandwidth but the price paid for this is increased network latency due to route discovery process.

ii. Proactive Routing Protocols

Proactive routing protocols are also called as table driven routing protocols. In this every node maintain routing table which contains information about the network topology even without requiring it. This feature although useful for datagram traffic, incurs substantial signalling traffic and power consumption [2]. The routing tables are updated periodically whenever the network topology changes. Proactive protocols are not suitable for large networks as they need to maintain node entries for each and every node in the routing table of every node [3]. These protocols maintain different number of routing tables varying from protocol to protocol.

There are various well known proactive routing protocols. Example: DSDV, OLSR, WRP etc. *Destination sequenced distance vector (DSDV) routing protocol* is table driven protocol based on the Distributed Bellman Ford Algorithm. The improvements made to the Bellman Ford algorithm include the freedom from loops in routing tables [2]. In this each node maintain routing table which contains next hop, number of hops to reach the destination, sequence number. Each node appends its. DSDV has large overhead due to routing tables.*WRP (wireless routing protocol)* is enhanced version of DSDV. Being proactive protocol it maintains routing information in the routing table. There are four types of tables maintained in this protocol namely distance table, routing table, link cost table, message retransmission list.

Optimised link state routing (OLSR) is based on the link state algorithm. OLSR protocol performs hop by hop routing i.e. each node uses its most recent information to route a packet [5].In this, MPR (Multipoint Relay nodes) are selected based on the greedy algorithm. The source node select nodes as MPR which are at one hop away from it and are able to cover the whole network.MPR are used to diffuse control message in the network which helps to reduce overhead. Whole network is covered through these MPR shown in Figure.4. Basic idea behind the MPR in the network is to reduce flooding in the network. The source node communicates with its two-hop neighbours through these MPR. The source node pass the control message to its MPR and the nodes which are not the MPR but are only one-hop neighbours just process the messages without forwarding them.

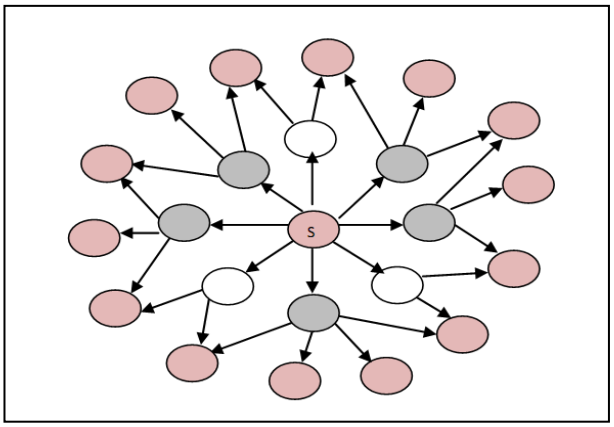


Fig. 4 : MPR structure

The source node S selects MPR from its one hop neighbours. The grey nodes represent MPR and white nodes are one hop neighbours but not the MPR nodes. The other nodes are two hop neighbours. The source node communicates with the two hop neighbours through its MPR.

Proactive protocols also has various advantages and disadvantages, being table driven protocols they increase the control messages in the network due which message overhead in the network increases .But at the same time due to routing information already present latency is reduced in the network. Proactive approaches also suffer from either out of date states or flooding of periodic updates [4].

iii. Hybrid Routing Protocol

While most of the protocols presented for MANET are either proactive or reactive protocols. There is a trade-off between proactive and reactive protocols. Proactive protocols have large overhead and less latency while reactive protocols have less overhead and more latency. So a Hybrid protocol is presented to overcome the shortcomings of both proactive and reactive routing protocols. Hybrid routing protocol is combination of both proactive and reactive routing protocol. It uses the route discovery mechanism of reactive protocol and the table maintenance mechanism of proactive protocol so as to avoid latency and overhead problems in the network. Hybrid protocol is suitable for large networks where large numbers of nodes are present. In this large network is divided into set of zones where routing inside the zone is performed by using reactive approach and outside the zone routing is done using reactive approach. There are various popular hybrid routing protocols for MANET like ZRP, SHRP,

ZRP (Zone Routing Protocol)[6] shown in Figure.5 uses the hybrid approach to routing. It is based on the merits of both proactive and reactive routing protocol. The nodes of a zone are divided into peripheral nodes and

interior nodes [7]. Every node in the network has a zone associated to it. The zone of a node is defined as the collection of nodes whose minimum distance from the node is not greater than the radius of the node. The minimum distance is defined in terms of number of hops from that node. The routing inside the zone i.e. intra-zone is done by using proactive approach. For intra-zone routing a node must know about its neighbours. The neighbours of nodes are defined as the nodes which are one hop away from particular node. The neighbour discovery is done by neighbour discovery protocol (NDP) so as to proactively monitor the network for intra-zone routing. The central node selects its zone by considering set of nodes whose distance

from the central node is not greater than the radius of the zone. These set of nodes are known as peripheral nodes.

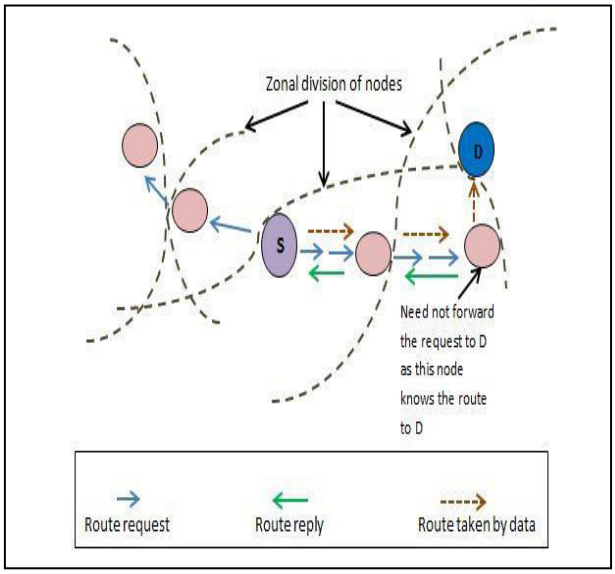


Fig. 5 : ZRP protocol

The intra-zone routing is done by intra-zone routing protocol (IARP).The IARP proactively monitors the network and maintains routes inside the zone. Outside the zone route discovering based on reactive approach is done to maintain routes. The Inter zone routing protocol (IERP) is responsible for maintaining the routes. Route discovery is done through a process called boarder casting. It is a packet delivery process through which nodes deliver packets to their peripheral nodes. In the route discovery mechanism source nodes initiate the route discovery it first checks whether destination is inside the zone or outside it, if it is inside the zone then the route is already available in the source node otherwise it send the query packet to its peripheral nodes, these nodes then verify whether the destination is inside their zone or not. In this way route discovery is been done.

IV. COMPARISON OF PROTOCOLS

The comparison among the different types of routing protocols is shown in Table.1.

Table.1 Parametric Comparison

Parameters	Reactive protocol	Proactive protocol	Hybrid protocol
Routing philosophy	Flat	Flat/Hierarchical	Hierarchical
Routing scheme	On demand	Table driven	Combination of both
Routing overhead	Low	High	Medium
Latency	High due to flooding	Low due to routing tables	Inside zone low outside similar to Reactive protocols
Scalability level	Not suitable for large networks	Low	Designed for large networks
Availability of routing information	Available when required	Always available stored in tables	Combination of both
Periodic updates	Not needed as route available on demand	Yes. Whenever the topology of the network changes	Yes needed inside the zone
Storage capacity	Low generally Depends upon the number of routes	High ,due to the routing tables	Depends on the size of Zone, insid0e the zone sometimes high as proactive protocol
Mobility support	Route maintenance	Periodical updates	Combination of both

Summary of protocols on the basis of advantages and disadvantages is shown in Table.2.

Table.2 Pros and Cons Comparison

Protocol	Advantages	Disadvantages
Proactive	Information is always available. Latency is reduced in the network	Overhead is high, Routing information is flooded in the whole network

Reactive	Path available when needed overhead is low and free from loops.	Latency is increased in the network
Hybrid	Suitable for large networks and up to date information available	Complexity increases

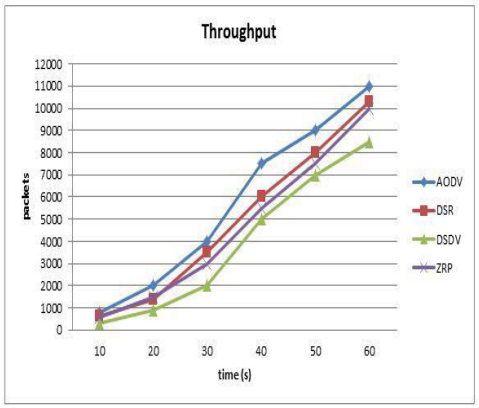
V. RESULT

Due to dynamic topology of ad hoc networks routing is one of the challenging issues in these networks. There are various types of routing protocols and these are suitable for different situations. It is seen that due to route discovery mechanism by reactive routing protocols overhead is very low in these protocols in contrast to proactive routing protocol in which overhead increases due to routing information stored in routing tables. But due to route discovery process the latency in the Reactive protocols increases whereas latency is very low in proactive protocols due to the fact that the routing information is already being stored in routing table and is available whenever needed. The Hybrid protocols have combined the advantages of both Reactive and Proactive protocols. The latency is decreased by using proactive protocol inside the zone and overhead is decreased by using reactive protocol outside the zone. Hence a protocol is presented which improves the performance of network by using the advantages of both reactive and proactive protocols.

A. Performance Metrics

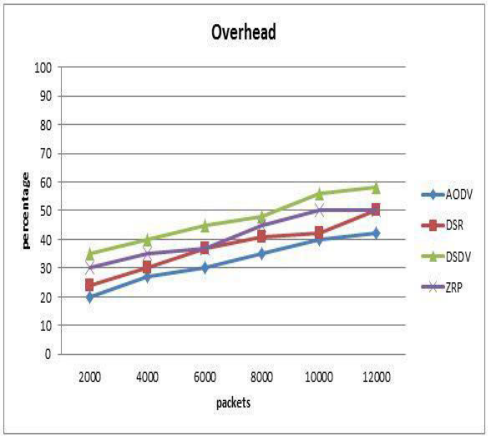
Throughput: This is the parameter related to the channel capacity. It is defined as the maximum possible delivery of the messages over the channel. It is usually measured in bits per second. The result is shown in Figure.6.

Fig. 6 : Result 1(Throughput)



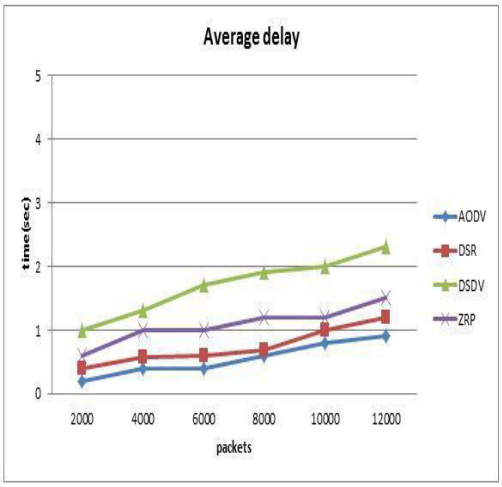
Routing Overhead: It is defined in terms of number of control packets need to be sent for the route discovery as well as route maintenance so as to send data packets. The result is shown in Figure.7.

Fig. 7 : Result 2 (overhead)



Average delay: it is defined as the time taken by the packet to reach from source to destination. It is measured in seconds. It is also known as end to end delay. The result is shown in Figure.8.

Fig. 8 : Result 3 (Average delay)



Packet delivery ratio: It is defined as the ratio of incoming data packets to the received data packets. We can understand that AODV has the better packet delivery ratio from the result of throughput shown in Figure.6.

Scalability: It is defined as the performance of routing protocols in presence of large number of nodes. Generally the performances of routing protocols degrade in presence of large number of nodes. We can compare this metric among the routing protocols and can say that AODV is the most scalable of all the

routing protocol, all other metrics regarding this protocol is better than the others.

VI. CONCLUSION

In this paper an effort has been made on the comparative study of Reactive, Proactive and Hybrid routing protocols. A comparison of three protocols has been presented in the form of table. Various advantages and disadvantages of these protocols are also presented in the form of table. There are various shortcomings in different routing protocols and it is difficult to choose routing protocol for different situations as there is trade-off between various protocols. The field of mobile ad-hoc networks is very vast and there are various challenges that need to be met, so these networks are going to have widespread use in the future.

REFERENCES

[1] Tarek Sheltami and Hussein Mouftah "Comparative study of on demand and Cluster Based Routing protocols in MANETs", IEEE conference, pp. 291-295, 2003.

[2] Elizabeth M. Royer"A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks" University of California, Santa Barbara Chai-Keong Toh, Georgia Institute of Technology, IEEE Personal Communications, pp. 46-55, April 1999.

[3] Krishna Gorantala , "Routing Protocols in Mobile Ad-hoc Networks", A Master' thesis in computer science, pp-1-36, 2006.

[4] Abdellah Jameli, Najib Naja and Driss El Ouadgiri "Comparative Analysis of Ad Hoc Networks Routing Protocols For Multimedia Streaming", IEEE, 1999.

[5] Shakkeera "Optimal path selection technique for Flooding in Link State Routing Protocol Using Forwarding Mechanisms in MANET".

[6] Zygmunt J. Haas, senior member IEEE and Marc R. Pearlman, member, IEEE "The performance of query control schemes for the zone routing protocol" ieee/acm transactions on networking, vol. 9, no. 4, august 2001

[7] Nicklas Beijar "Zone routing protocol" Networking Laboratory, Helsinki University of Technology, P.O. Box 3000, FIN-02015 HUT, Finland".

◈◈◈

Load Balancing in Computational Grids Using Ant Colony Optimization Algorithm

Sowmya Suryadevera[1], Jaishri Chourasia[2], Sonam Rathore[3] & Abdul Jhummarwala[4]

[1,2&3]MITS,Lakshmangarh, [4]Bisag,Gandhinagar

Abstract – Grid computing is the combination of computer resources from multiple administrative domains for a common goal. Load balancing is one of the critical issues that must be considered in managing a grid computing environment. It is complicated due to the distributed and heterogeneous nature of the resources. An Ant Colony Optimization algorithm for load balancing in grid computing is proposed which will determine the best resource to be allocated to the jobs, based on resource capacity and at the same time balance the load of entire resources on grid. The main objective is to achieve high throughput and thus increase the performance in grid environment.

Keywords - *Grid computing, Ant colony optimization, Grid load balancing, performance, Grid, ANT algorithm, Scheduling.*

I. INTRODUCTION

Grid computing offers seamless access to rare and limited resources. Grid computing (or the use of a computational grid) is applying the resources of many computers in a network to a single problem at the same time usually to solve a scientific or technical problem that requires a great number of computer processing cycles or access to large amounts of data. A computational grid is the cooperation of distributed computer systems where user jobs can be executed on local or remote computer systems. On one side it provides the user with access to locally unavailable resource types on the other hand there is the expectation that a larger number of resources are available. A data grid denotes systems that provide a hardware and software infrastructure for synthesizing new information from data repositories that are distributed in a wide area network. Load balancing [1] and resource scheduling algorithm plays important role in the resource management and also gives much impact in the performance point of view.

Scheduling is process that maps and manages the execution of inter- dependent tasks on the distributed resources. The demand for scheduling is to achieve high performance computing. It aims to find a suitable allocation of resources for each job. It allocates suitable resources to tasks so that the execution can be completed to satisfy objective functions imposed by users. In Grid environments, scheduling decisions must be made in the shortest time possible, because there are many users competing for resources, and time slots desired by one user could be taken up by another user at any moment.

The load balancing mechanism aims to equally spread the load on each computing node, maximizing their utilization and minimizing the total task execution time. Load balancing is a mechanism which equally spreads the load. It minimizes the response time and improves the resource utilization rate. Load balancing in grid can be done based on the conditions of the system such as static or dynamic [7].

Static load balancing:

In static load balancing, algorithms scheduling is based on a default policy. Condition of system in every scheduling will be specified and on that base scheduling takes place and no more scheduling will take place until the work is done.

Dynamic load balancing:

A dynamic load balancing algorithm adapts its decision with the system that means processing duties with changing system condition. Dynamic load balancing makes decision on basis of the present system condition and quickly adapts with workload fluctuations.

This paper presents algorithm based on the general ant adaptive scheduling heuristics and an added in load balancing guide component. Section 2 describes the use variants of ant colony optimization (ACO) algorithm in grid computing. The proposed algorithm is discussed in Section 3.The concluding remarks are highlighted in Section 4.

II. ACO ALGORITHMS IN GRID

ACO [1] is inspired by a colony of ants that work together in foraging behaviour. This behaviour encouraged ants to find the shortest path between their nest and food source. Every ant will deposit a chemical substance called pheromone on the ground after they move from the nest to food sources and vice versa. Therefore, they will choose an optimal path based on the pheromone value. The path with high pheromone value is shorter than the path with low pheromone value. This behaviour is the basis for a cooperative communication. There are various types of ACO algorithm such as Ant Colony System (ACS), Max- Min Ant System (MMAS), Rank-Based Ant System (RAS) and Elitist Ant System (EAS) [5].

ACO has been applied in solving many problems in scheduling such as Job Shop Problem, Open Shop Problem, Permutation Flow Shop Problem, Single Machine Total Tardiness Problem, Single Machine Total Weighted Tardiness Problem, Resource Constraints Project Scheduling Problem, Group Shop Problem and Single Machine Total Tardiness Problem with Sequence Dependent Setup Times [5]. A recent approach of ACO researches in the use of ACO for scheduling job in grid computing [6]. ACO algorithm is used in grid computing because it is easily adapted to solve both static and dynamic combinatorial optimization problems. In [2], ACO has been used as an effective algorithm in solving the load balancing problem in grid computing. The process taken by ACO will consider the pheromone value which depends on the time taken by each resource to process jobs. It does not consider the capacity of resources such as their bandwidth, processor speed and load.

In [4], two distributed artificial life-inspired load balancing algorithm are introduced, which are ACO and Particle Swarm Optimization (PSO). In the proposed algorithm, an ant acts as a broker to find the best node in term of the pheromone value stored in the pheromone table. The node with the lightest load is selected as the best node. The position of each node in the flock can be determined by its load in PSO. The particle will compare the load of nodes with its neighbours and will move towards the best neighbour by sending assigned jobs to it. The proposed algorithm performed better than ACO for job scheduling where jobs are being submitted from different sources and different time intervals. However, PSO uses more bandwidth and communication compared to ACO.

A study in [3] proposed a new algorithm that is based on an echo intelligent system, autonomous and cooperative ants. In this proposed algorithm, the ants can procreate and also can commit suicide depending on existing condition. Ant level load balancing is proposed to improve the performance of the mechanism. Ants are created on demand during their lives adaptively to achieve the grid load balancing. The ants may bear offspring when they detect the system is drastically unbalanced and commit suicide when they detect equilibrium in the environment. The ants will care for every node visited during their steps and record node specifications for future decision making. Theoretical and simulation results indicate that this new algorithm surpasses its predecessor. However, the pheromone values were not updated in this proposed algorithm which enables the assignment of jobs to the same resource.

ACO algorithm for load balancing in distributed systems through the use of multiple ant colonies is proposed in [8]. In this algorithm, information on resources is dynamically updated at each ant movement. Load balancing system is based on multiple ant colonies information. Multiple ant colonies have been adopted such that each node will send a coloured colony throughout the network. Coloured ant colonies are used to prevent ants of the same nest from following the same route and also enforcing them to be distributed all over the nodes in the system and each ant acts like a mobile agent which carries newly updated load balancing information to the next nodes. This proposed algorithm has been compared with the work-stealing approach for load balancing in grid computing. Experimental result shows that multiple ant colonies work better than work-stealing algorithm in term of their efficiency. However, the multiple ant colonies do not consider resources capacity and jobs characteristics. This can make matching the jobs with the best resources a difficult task for the scheduling algorithm.

An enhanced ant algorithm for task scheduling in grid computing was proposed in [9], which gives better throughput with a controlled cost. The proposed scheduling algorithm increased the performance in terms of low processing time and low processing cost when applied to a grid application with a large number of jobs such as parameter sweeps application. This algorithm works effectively in minimizing the processing time and processing cost of the jobs. The simulation results of various scheduling algorithm such as modified ant algorithm and cost controlled algorithm are also compared. The result shows that this enhanced algorithm works better than the ant algorithm. By considering the processing cost, this enhanced ant algorithm is more suitable for wide use. However, this algorithm does not consider the size of the jobs which leads to inappropriate assignment of jobs to resources.

Based on the previous research discussed above, there have been some or the other problems while scheduling or balancing load in a grid environment. Taking some of

the problems into consideration ant colony optimization algorithm for load balancing in grid computing is proposed which will determine the best resource to be allocated to the jobs, based on resource capacity and at the same time balance the load of entire resources on grid.

III. PROPOSED ANT ALGORITHM FOR GRID LOAD BALANCING

When a resource j enrolls into the grid system, it is asked to submit its performance parameters, such as the number of processor, processing capability MIPS (Mega Instruction per Second) of each processor, its RAM capability and the communication ability etc. Based on these parameters the initial pheromone value i.e. trail intensity for resource j is calculated as:

$$\tau j(0) = n \times p + Sj + M \quad (1)$$

where n is the No. of processors, p is MIPS of each processor, the parameter Sjis the communication bandwidth ability and M is the RAM capability of resource j.

The trail intensity at time t is updated and given by:

$$\tau j(t) = \rho \times \tau j(t-1) + \Delta\tau j(t-1,t) \quad (2)$$

Where, $\Delta\tau j(t-1,t)$ is the variety of quantity of the trail substance laid on the path from the scheduling center to resource j between time t-1 and t; ρ is the permanence of pheromone ($0 < \rho < 1$); $(1-\rho)$ is an evaporation of pheromone.

When job is assigned to resource j the pheromone value of that resource is updated such that $\Delta\tau j(t-1,t) = -k$, where k is the coefficient relevant to computation workload and communication quantity of the job.

When a job is successfully completed and the resource j is released then there would be an increment in the pheromone value of that resource, $\Delta\tau j(t-1,t) = Ce \times k$, where Ce is the encouragement coefficient.

When a job fails and the resource j is released then there would be decrement in the pheromone value of the resource, $\Delta\tau j(t-1,t) = Cp \times k$, where Cp is the punishment coefficient.

Different chosen values of the above mentioned coefficients ρ, k, Ce, Cp will change the value of $\Delta\tau j(t-1,t)$, and when to update $\tau j(t)$ cause different instantiation of the ant algorithm.

The possibility of task assignment to every resource will be recomputed as:

$$\rho_j(t) = \frac{[\tau_j(t)]^{\alpha} * [\eta_j(t)]^{\beta}}{\sum_u [\tau_u(t)]^{\alpha} * [\eta_u]^{\beta}} \quad (3)$$

where $\tau j(t)$ the trail intensity on the path form scheduling center to resource j at time t; ηj is called visibility, namely the innate performance quantity of the resource j ($\tau j(0)$); α is the parameter on the relative importance of trail intensity; β is the parameter on the relative importance of visibility.

The jobs are allocated to resources based on the pheromone value calculated by considering resource parameters. But by the above method if a particular resource always completes jobs successfully then the resource gets loaded heavily. This results in the bottleneck in the grid and influences to completing the jobs. Therefore we introduce load balancing factor in the ant algorithm to improve the load balancing capability.

We thus introduce the load balancing factor λjof resource j, which would depend on the job finishing rate of the resource j. The load balancing factor will change the trial intensity from $\Delta\tau j$to $\Delta\tau j + C\ \lambda j$ ($C > 0$ is a coefficient of the load balancing factor), the more jobs finished the more increases the trial intensity,contrarily, the more jobs not completed, the more decreases the trail intensity. So by introducing the load balancing factor the load on all the resources in the grid will be balanced.

IV. CONCLUSION AND FUTURE WORK

The proposed algorithm is expected to determine the best resource to be allocated to the jobs, based on resource capacity and at the same time to balance the load in grid. The algorithm considers various resource parameters (MIPS, communication bandwidth, No. of processors and memory) for calculating pheromone i.e. the resource capability for executing various jobs thus allocating the best resource for the job and at the same time balance the load of all the resources. The algorithm is expected to achieve better throughput and hence increase the overall performance in the grid environment. In the future the algorithm will be implemented using GridSim simulator.

VI. ACKNOWLEDGMENTS

We would like to thank Mr. P. K. Bishnoi of MITS, Lakshmangarh and Mr. M.B. Potdar of BISAG, Gandhinagar for their support and guidance.

REFERENCES

[1] Jagdish Chandra Patni, Dr. M.S.Aswal, Om Prakash Pal and Ashish Gupta, "Load balancing Strategies for Grid Computing" presented at 3rd international conference on electronics computer technology (ICECT), vol.3, pp. 239-243, 2011.

[2] Y. Li, "A Bio-inspired Adaptive Job Scheduling Mechanism on a Computational Grid," International Journal of Computer Scienceand Network Security (IJCSNS), vol. 6(3), pp. 1-7, 2006.

[3] M. Salehi and H. Deldari, "Grid load balancing using an echo system of intelligent ants," presented at Proceedings of the 24th IASTED international conference on Parallel and distributed computing and networks, 2006.

[4] A. Moallem, and S. A. Ludwig, "Using Artificial Life Techniques for Distributed Grid Job Scheduling" presented at ACM Symposium on Applied Computing (SAC 2009), Hawaii, U.S.A., pp. 1091 – 1097, 2009.

[5] M. Dorigo and T. Stützle, Ant colony optimization, Cambridge,Massachusetts, London, England: MIT Press, 2004.

[6] S. Fidanova and M. Durchova, "Ant algorithm for grid scheduling problem," Lecture Notes in Computer Science, vol. 3743, pp. 405- 412, 2006.

[7] Moradi, M.Dezfuli, M.A.Safavi, "A new time optimizing probabilistic load balancing algorithm in grid computing" presented at 2nd International Conference on Computer Engineering and Technology (ICCET), vol. 1, pp. v1232-v1237, 2010.

[8] A. Ali, M. A. Belal and M. B. Al-Zoubi, "Load Balancing of Distributed system Based on Multiple Ant Colonies Optimization,"American Journal of Applied Sciences, vol. 7(3), pp. 433-438, 2010.

[9] K. Sathish and A. Reddy, "Enhanced ANT Algorithm Based Load Balanced Task Scheduling in GRID Computing," IJCSNS, vol. 8, pp. 219, 2008.

◈◈◈

First Hop Security For IPv6

Smriti Joshi[1] & Pushpendra Tyagil[2]

[1]Amity university, Noida, India , [2]Shobhit University, Meerut, India

Abstract - There are a growing number of large-scale IPv6 deployments at enterprise, university, and government networks. For the success of each of these networks, it is important that the IPv6 deployments are secure and are of a service quality that equals that of the existing IPv4 infrastructure. Network users have an expectation that there is functional parity between IPv4 and IPv6 and that on each of these protocols security and serviceability concerns are similar. From the network operator perspective there is a similar assumption that both IPv4 and IPv6 are secure environments with a high degree of traceability and quality assurance. This paper provides an introduction to the concept of first-hop security and discusses the challenge of trying to apply the techniques currently enforced in IPv4 networks to emerging IPv6 deployments.

Keywords - *First Hop,Hop Security,IPv6 network security.*

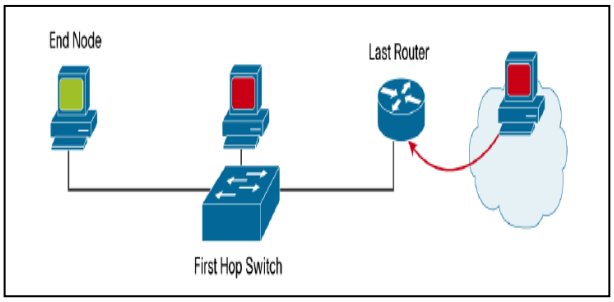
I. INTRODUCTION

The term "first-hop security" is generally used to refer to the security policies and mechanisms that can be locally employed on a network segment to protect it from attack. There are different points in an enterprise network (local or wide area network) where first-hop security can be enforced:

- the end nodes
- the first-hop switch
- the first-hop router

Enforcing first-hop security at the end-nodes leads to a distributed security model in which each node protects itself. The upside of this model is that it does not result in single points of failure, where the failure of a single network element affects a large portion of the network. This security model also does not increase network complexity because security is enforced by the hosts rather than network elements like the first-hop switches or routers. The downsides are that this distributed security model makes the network generally harder to manage than one based on a centralized model, and vulnerabilities are nearly impossible to mitigate at the end nodes.

Enforcing first-hop security at the first-hop switch leads to a **centralized security model** that puts the first-hop switch in charge of protecting the end nodes. The upside of this centralized model is that it tends to be easier to manage. Also, the first-hop switch becomes a natural point at which to mitigate a number of common network attacks. The downsides are that it requires increased intelligence/complexity in the first-hop switch and it clearly introduces a single point of failure; if the first-hop switch is compromised, all protection is gone.



Finally, first-hop security can be enforced at the first-hop router. The upside of this model is that it leads to a **more protected centralized model** that is simple to manage and can protect all subnet elements from external attack. On the other hand, this model cannot protect against attacks originated from local nodes.

Enforcing first-hop security usually mitigates a number of attacks. They include -- but are not limited to -- the following:

- Address spoofing attacks
- Layer 2/Layer 3 interface attacks, like address-resolution
- Denial of Service (DoS) attacks, based on either Layer 2 or Layer 3

II. FIRST-HOP SECURITY IN IPv4

In IPv4, the Address Resolution Protocol (ARP) is used to map IPv4 addresses into link-layer addresses. ARP runs directly on top of Ethernet and is a very simple protocol with fixed-length packets and no options or extensions. As a result, ARP traffic can be easily monitored to detect ARP spoofing attacks. Tools such as arpwatch have been readily available for a long

time to monitor ARP traffic on enterprise networks. Additionally, IPv4 switches can reliably block spoofed ARP traffic.

In IPv4 networks, automatic network configuration typically employs the Dynamic Host Configuration Protocol (DHCP), which can be exploited to perform man in the middle or DoS attacks. In order to mitigate such attacks, some IPv4 switches implement a "DHCP snooping" functionality, so that outgoing DHCP-server packets are only allowed on specific ports. This type of functionality effectively and simply mitigates DHCP-based attacks by blocking the attack packets at the local switch before hackers reach the victim hosts.

IPv4 networks may also be subject to address-spoofing attacks, where a host may try to impersonate either another host on the same network segment or a host on a remote network. Since each IPv4 host is typically assigned a single address, the scale of the spoofing problem is very limited; a first-hop switch can mitigate this problem by allowing only a few addresses per port.

Finally, it is generally desirable to track address usage in a network -- that is, to keep a log of which node used which address(es) at which point in time. This is useful for correlating node activities, such as identifying malware-infected systems in a network. It takes fewer resources to track IPv4 addresses than IPv6 addresses; since addresses are typically assigned by a DHCP server, the aforementioned server is an obvious and straightforward choice for logging such information.

III. FIRST-HOP SECURITY IN IPv6

A key difference between IPv6 and IPv4 subnets is that IPv6 subnets are typically assigned a much larger address space (usually a /64). Because IPv6 can accommodate a larger number of nodes, all relevant network elements must also be prepared to handle such a large number of nodes/addresses. Additionally, IPv6 nodes are typically assigned more than one IPv6 address. At the very least, they are assigned a link-local unicast address (which may look something like this: fe80::/10) and one global unicast address.

Finally, a number of operating systems (notably Windows Vista and Windows 7) support and enable by default some flavor of "temporary addresses" (usually referred to as "privacy addresses"), which are configured in addition to the traditional auto-configured addresses. These privacy addresses are short-lived and recycled over time, which means that not only do IPv6 hosts employ more than one address, but the set of employed addresses varies over time.

In the IPv6 world, ARP has been replaced by the "Neighbor Discovery" (ND) protocol. ND employs Internet Control Message Protocol version 6 (ICMPv6) messages, so it runs on top of IPv6. When compared with IPv4's ARP, ND provides increased flexibility. However, it also results in increased complexity: In particular, since it runs on top of IPv6, ND messages can potentially include IPv6 extension headers and may be fragmented into multiple IPv6 fragments.

IV. ICMPv6 AND NEIGHBOUR DISCOVERY PROTOCOL

This section introduces the Internet Control Message Protocol Version 6 (ICMPv6). In comparison with IPv4, IPv6 has an increased set of capabilities to simplify end-system auto configuration while at the same time running service detection by means of ICMP. Because of these new ICMP capabilities, the importance of ICMP for IPv6 is much higher than it ever was for IPv4. One of the new functionalities within ICMPv6 is the Neighbor Discovery Protocol, which in its base specification is a non-authenticated protocol. NDP is an application and operates above ICMPv6. NDP makes heavy usage of multicast packets for on-the-wire efficiency.

The functional applications of NDP include:

- Router discovery
- Autoconfiguration of addresses (stateless address autoconfiguration [SLAAC])
- IPv6 address resolution (replaces Address Resolution Protocol [ARP])
- Neighbor reachability (neighbor unreachability detection [NUD])
- Duplicate address detection (DAD)
- Redirection

V. WHAT IS SeND?

Secure Neighbor Discovery is a protocol that enhances NDP with three additional capabilities:

- **Address ownership proof**
  - Makes stealing IPv6 addresses “impossible”
  - Used in router discovery, DAD, and address resolution
  - Based upon Cryptographically Generated Addresses (CGAs)
    - Alternatively also provides non-CGAs with certificates
- **Message protection**
  - Message integrity protection
  - Replay protection

- Request/response correlation
- Used in all NDP messages

- **Router authorization**
  - Authorizes routers to act as default gateways
  - Specifies prefixes that routers are authorized to announce "on-link"

While SeND provides a significant uplift to the IPv6 neighbor discovery technology by introducing the above enhancements, it does not, for example, provide any end-to-end security and provides no confidentiality.

It is important to understand that SeND is **not** a new protocol and still remains a protocol operating on the link. Secure Neighbor Discovery is just an "extension" to NDP and defines a set of new attributes:

- **New network discovery options**

CGA, Nonce, Timestamp, and RSA

NONCE: In order to prevent replay attacks, two new neighbor discovery options, Timestamp and Nonce (a random number), are introduced. Given that neighbor and router discovery messages are in some cases sent to multicast addresses, the Timestamp option offers replay protection without any previously established state or sequence numbers. When the messages are used in solicitation-advertisement pairs, they are protected with the Nonce option.

Purpose: These options provide a security shield against address theft and replay attacks.

- **New network discovery messages**

CPS (Certificate Path Solicitation), CPA (Certificate Path Advertisement)

CPA: Certification paths, anchored on trusted parties, are expected to certify the authority of routers. A host must be configured with a trust anchor to which the router has a certification path before the host can adopt the router as its default router. Certification path solicitation and advertisement messages are used to discover a certification path to the trust anchor without requiring the actual router discovery messages to carry lengthy certification paths. The receipt of a protected router advertisement message for which no certification path is available triggers the authorization delegation discovery process.

Purpose: Identifying valid and authorized IPv6 routers and IPv6 prefixes of the network segment. These two messages complement the already existing NDP messages (NS, NA, RA, RS, and Redirect).

- **New rules**

Purpose: These rules describe the preferred behavior when a SeND node receives a message supported by

SeND or not supported by SeND.

SeND technology works by having a pair of private and public keys for each IPv6 node in combination with the new options (CGA, Nonce, Timestamp, and RSA). Nodes that are using SeND cannot choose their own interface identifier because the interface identifier is cryptographically generated based upon the current IPv6 network prefix and the "public" key. However, the CGA interface identifier alone is not sufficient to guarantee that the CGA address is used by the appropriate node.

For this purpose SeND messages are signed by usage of the RSA public and private key pair. For example if node 1 wants to know the MAC address of node 2, it will traditionally send a neighbor solicitation request to the node 2 solicited node multicast address. Node 2 will respond with a corresponding neighbor advertisement containing the MAC address to IPv6 address mapping. Node 2 will in addition add the CGA parameters (which include among others the public key) and a private key signature of all neighbor advertisement fields. When node 1 receives this neighbor advertisement it uses the public key to verify with the CGA address the private key signature of node 2. Once this last step has been successfully completed, the binding on node 1 of the MAC address and CGA address of node 2 can be successfully finalized.

VI. SeND DEPLOYMENT CHALLENGES

While the construction of a CGA address is a rather lightweight action because it "only" requires hosts to be cryptocapable (generate RSA key pairs, RSA sign NDP messages, and RSA verify messages). On the other hand, the SeND capability for router authorization is a much more heavyweight technology because it relies upon Certificate Authority (CA) implementation for hosts to trust routers. Also for routers to be trusted, they need some PKI implementation so that they can get a certificate from the CA and for obtaining and maintaining the certificate chain in case of hierarchical CAs. It is a pragmatic assumption that many hosts will not be deployed with CA certificates due to the complexity involved. Another challenge to deploy SeND is the bootstrapping of the trust relationship. To access the Certificate Revoke List (CRL) and the time server, the host would need to access these devices through a router it does not trust yet. A way to work around this challenge is to preprovison the host with certificates and ship them to users.

VII. SECURING AT THE FIRST HOP

The first hop for an end node is very often a Layer 2 switch. By implementing the right set of security features this switch has a potential to solve many of the caveats attached to a SeND deployment and increase the link security model. The first hop switch is strategically located to learn about all its neighbors, and hence the switch can easily either allow or deny certain types of traffic, end-node roles, and claims. In its central position, the first hop switch can fulfill a number of functions. It will inspect the ND traffic and provide information about Layer 2/Layer 3 binding and monitor the use of ND by host to spot potentially abnormal behaviors. Ultimately, the switch can block undesired traffic such as rogue Router Advertisement (RA), rogue DHCP server advertisement, and data traffic coming from undesired IP addresses or prefixes.

VIII. CONCLUSION

First-hop security aims at improving the security of a local network by employing a number of mitigation techniques. While feature parity between IPv6 and IPv4 is highly desirable, some of the differences between these two protocols can make achieving it difficult. The standards and vendor communities are working to overcome these issues and bring the well-known IPv4 mitigations to the IPv6 world. Aside from the ongoing research and development work, IPv6 awareness needs to be raised among network and security administrators, so that differences between old and new Internet protocols do not negatively impact existing and emerging IPv6 deployments.

REFERENCES

[1] Davies, J. (2003). Understanding IPv6. Redmond, WA: Microsoft Press.

[2] White paper (2010) on IPv6 First Hop Security—Protecting Your IPv6 Access Network by Cisco.

[3] Green, D., & Grillo, B. (2005), The State of IPv6: a Department of Defense perspective. DoD IPv6 Standards Working Group, Retrieved June 12, 2005 from http://ipv6.disa.mil/docs/State-of-IPv6-Final-7Feb05.pdf.

[4] Cheswick, W., Bellovin, M., & Rubin, A. (2003). Firewalls and Internet security (2nd ed.). Boston: Addison-Wesley. Cisco Systems, (2005). Implementing security for IPv6. Cisco Systems, Retrieved July5,2005from http://www.cisco.com/en/US/products/sw/iosswrel/ps5187/products_configuration_guide_chapter09186a00801d65f4.html

[5] Article on First-hop security in IPv6 by searchenterpriseWAN.com found at online at http://searchenterprisewan.techtarget.com/tip/First-hop-security-in-IPv6

◈◈◈

Improve Data Quality in Sales Management Using Association Rule in Multi-Relational Data Mining

Sunil Kumar & Pravin Kumar

School of Information and Communication Technology, Gautam Buddha University,
Greater Noida, Uttar Pradesh, India

Abstract - In this paper, we improve the data quality and increase the sales product using the association rules in multi relational data mining. This problem arise from reduce sales and decrease the accuracy of data. And also increase the speed of processing in different reason. In multi relational data mining reduced the mistake in database. The probability of mistake identify by algorithm. We cannot identify mistake in single table. But also using multi relational data mining we identify mistake in relational table. Also discuss join the many different tables using association rules in database.

Keywords - *Data Miming; Multi-relational Data Mining; Association Rules; Data Quality.*

I. INTRODUCTION

Data mining technology has obtained extensive research, application and development since its beginning, the research field is very extensive. But data mining is based on common table structure in data organizing form of research [1]. Multi-relational Data Mining is inspired by the relational model. This model presents a number of techniques to store, manipulate and retrieve complex and structured data in a database consisting of a collection of tables. It has been the dominant paradigm for industrial database applications during the last decades, and it is at the core of all major commercial database systems, commonly known as relational database management systems (RDBMS). A relational database consists of a collection of named tables, often referred to as relations that individually behave as the single table that is the subject of Propositional Data Mining [2]. We will assume that the data to be analysed is stored in a relational database. A relational database consists of a set of tables and a set of associations between pairs of tables describing how records in one table relate to records in another table. Both tables and associations are also known as relations, so we will use the former terminology to be able to distinguish between the two concepts. An association between two tables describes the relationships between records in both tables. The nature of this relationship is characterised by the multiplicity of the association. The multiplicity of an association determines whether several records in one table relate to single or multiple records in the second table. Also, the multiplicity determines whether every record in one table needs to have at least one corresponding record in the second table [3]. Multi-Relational Data Mining (MRDM) deals with knowledge discovery from relational databases consisting of one or multiple tables. As a typical technique for MRDM, inductive logic programming (ILP) has the power of dealing with reasoning related to various data mining tasks in a "unified" way. Like granular computing (GrC), ILP-based MRDM models the data and the mining process on these data through intension and extension of concepts. Unlike GrC. However, the inference ability of ILP-based MRDM lies in the powerful PROLOG like search engine. Although this important feature suggests that through ILP, MRDM can contribute to the foundation of data mining (FDM), the interesting perspective of "ILP based MRDM for FDM" has not been investigated in the past [4].

II. DATA MINING

The primary ingredient of any Data Mining exercise is the database. A database is an organized and typically large collection of detailed facts concerning some domain in the outside world. The aim of Data Mining is to examine this database for regularities that may lead to a better understanding of the domain described by the database. In Data Mining we generally assume that the database consists of a collection of individuals [5].Depending on the domain, individuals can be anything from customers of a bank to molecular compounds or books in a library. For each individual, the database gives us detailed information concerning the different characteristics of the individual, such as the

name and address of a customer of a bank, or the accounts owned. When considering the descriptive information, we can select subsets of individuals on the basis of this information. For example we could identify the set of customers younger than 18. Such intentionally defined collections of individuals are referred to as subgroups [6].While considering different subgroups, we may notice that certain subgroups have characteristics that set them apart from other subgroups. For instance, the subgroup 'age under18' may have a negative balance on average. The discovery of such a subgroup will lead us to believe that there is a dependency between age and balance of a customer. Therefore, a methodical survey of potentially interesting subgroups will lead to the discovery of dependencies in the database. Clearly, a good definition of the nature of the dependency (e.g. deviating average balance) is essential to guide the search for interesting subgroups. Such a statistical definition is known as an interestingness measure or score function. Interesting subgroups are a powerful and common component of Data Mining, as they provide the interface between the actual data in the database and the higher-level dependencies describing the data. Some Data Mining algorithms are dedicated to the discovery of such interesting subgroups [7]. However, interesting subgroups are a limited means of capturing knowledge about the database, because by definition they only describe parts of the database. Most algorithms will therefore regard interesting subgroups not as the end product, but as mere building blocks for comprehensive descriptions of the existing regularities. The structures that are the aim of such algorithms are known as models, and the actual process of considering subgroups and laboriously constructing a complete picture of the data is therefore often referred to as modeling [8].We can think of the database as a collection of raw measurements concerning a particular domain. Each individual serves as an example of the rules that govern this domain. The model that is induced from the raw data is a concise representation of the workings of the domain, ignoring the details of individuals. Having a model allows us to reason about the domain, for example to find causes for diseases in genetic databases of patients. More importantly, Data Mining is often applied in order to derive predictive models. If we assume that the database under consideration is but a sample of a larger or growing population of individuals, we can use the induced model to predict the behavior of new individuals. Consider, for example, a sample of customers of a bank and how they responded to a certain offer. We can build a model describing how the response depends on different characteristics of the customers, with the aim of predicting how other customers will respond to the offer. A lot of time and effort can thus be saved by only approaching customers with a predicted interest [9]. Data mining has been an area looming just beyond statistical science for several years, and even an area that some statisticians evidently regard as overlapping with their territory. Yet many people may be unclear on how it differs from statistics applied to large or very large datasets, together to be sure with a lot of data management [10].

III. MULTI-RELATIONAL DATA MINING

Multi-Relational Data mining is inspired by the relational model. This model presents a number of techniques to store, manipulate and retrieve complex and structured data in a database consisting of a collection of tables. It has been the dominant paradigm for industrial database applications during the last decades, and it is at the core of all major commercial database systems, commonly known as relational database management systems (RDBMS) [11]. A relational database consists of a collection of named tables, often referred to as relations that individually behave as the single table that is the subject of Propositional Data Mining. Data structures more complex than a single record are implemented by relating pairs of tables through so-called foreign key relations. Such a relation specifies how certain columns in one table can be used to look up information in corresponding columns in the other table, thus relating sets of records in the two tables. Structured individuals (graphs) are represented in a relational database in a distributed fashion. Each part of the individual (node) appears as a single record in one of the tables. All parts of the same class for all individuals appear in the same table. By following the foreign keys (edges), different parts can be joined in order to reconstruct an individual. In our search for patterns in the relational database, we will need to query individuals for certain structural properties [12]. Relational database theory employs two popular languages for retrieving information from a relational database: relational algebra and the Structured Query Language (SQL). The former is primarily used in the theoretical settings, whereas the latter is primarily used in practical systems. SQL is supported by all major RDBMS. We employ an additional (graphical) language that selects individuals on the basis of structural properties of the graphs. This language translates easily into SQL, but is preferable because manipulation of structural expressions is more intuitive [13].

IV. MULTI-RELATIONAL ASSOCIATION MINING

Association rules mining is identified as one of the important problems in data mining. Let us first define the problem for a Database D containing a set of transactions, where each transaction contains a set of items [14]. The first mining technique to find

associations in multi relational data was Warmer. Warmer is a first order upgrade of Apriori [15]. WARMR is a recent relational association rule approach, which is a powerful ILP algorithm. WARMR discovers frequent Prolog queries that succeed with respect to a sufficient number of examples. The Prolog formulation is very general, as it allows the use of variables and multiple relations in the pattern. The flexibility of WARMR is a strong advantage over previous algorithms for the discovery of frequent patterns [16].Extended traditional associations to include association rules of forms A⇒¬B, ¬A⇒B, and ¬A⇒¬B in mono-database. These forms are called association rules which indicate associations between item sets [17]. However, we often need to obtain association rules across multiple databases. For example, facing different medical databases coming from different areas, the Center for Disease Control and Prevention is interested in the fact that which factors are relatively irrelevant or absolutely irrelevant although they arise frequently. Here, those absolutely irrelevant factors are important particularly in decision-making, which involve with mining strong association rules in multi database [18]. Multi-database association rule mining is a challenging and critical task since it requires knowledge of all the data stored in different locations and the ability to combine partial results into a single result from individual RDBMS.

V. DATA QUALITY

A data analysis seems important, the accuracy of the data in database is very essential as well because an analysis of right and correct data affects employers' decision making since their mistakes become fewer so that they can make more logical decisions, their benefits will increase compared to their competitors and their risk taking will be less. Since organizations and companies work based on data and their analyses, data and information are precious invests which play an important role in management decision making. Moreover, since data are processed in different stages, there will be problems in their analysis and collections. According to the statistics of data storage, 15-20% of data are inaccurate in one ordinary organization. In the meantime, low quality of data is the reason of this failure [19]. Data quality is defined based on the mentioned target so that it is not a term to be defined officially. In the literature, "suitable for application" or meeting users; needs is used. According to quality management, quality data is used to meet customers' needs. Despite general idea, quality doesn't mean zero mistakes [18].

In this way, we can improve data quality of input transaction with Algo.1 it is not necessary to extract all association rules of database. Extracting all association rules needs a lot of time and too much memory (time and space complexity). If we reduce the number of large itemsets, we can decrease a number of association rules. Therefore, we decrease time and space complexity. In addition, in Algo.1, we have just extracted association rules, which depend on input transaction and are adapted by one of the functional dependency. Therefore, we will reduce the number of large itemsets [19].

Algorithm 1 (Data Quality Mining)
Step1: Extract association rule, which depends on the input transaction (T) and is adapted by the functional dependency.
Step2: Separate compatible and incompatible rules.
Step3: Calculate the quality of input transaction.

In Algo.2, the number of large itemsets will be reduced because of a power set member must be subset of the large itemsets. Also In third step, there is a limitation for extracting association rules from the large itemsets. It causes to extract fewer association rules.

Algorithm 2(extracting association rules)
Step1: Extract power set of items in T
Step2: Extract the large itemsets (condition for each itemset: one of power set member must be subset of the large itemset)
Step 3: Extract association rules from the large itemset (condition: one of the functional dependencies must adapt Association rules).

VI. DATA QUALITY MINING EXPERIMENTS

To investigate the relationship between quality, number of tables, and the overall quality of table using association rules, we start by considering the case where for induction each repeatedly-tables example is assigned a single “integrated” table Yi, inferred from the individual Yij ’s by majority voting. For simplicity, and to avoid having to break ties, we assume that we always obtain an odd number of tables. The quality qi = Pr(Yi= Yj) of the integrated table Yi will be called the integrated quality. Where no confusion will arise, we will omit the subscript i for brevity and clarity.

(a). *Data Quality Functional Equation*

We first consider the case where all tables exhibit the same quality, that is, $pj = p$ for all j (we will relax this assumption later). Using $2N + 1$ tables with uniform quality p, the integrated tables quality q is:

$$q = \Pr(Yi = Yj) = \sum_{i=0}^{N} \binom{2N+1}{i} p^{2N+1} \cdot (1-p)^{i}$$

Which is the sum of the probabilities that we have more correct tables than incorrect (the index i corresponds to the number of incorrect tables). Not surprisingly, from the formula above, we can infer that the integrated quality q is greater than p only when $p > 0.5$. When $p < 0.5$, we have an adversarial setting where $q < p$, and, not surprisingly, the quality decreases as we increase the number of tables. Figure 1 demonstrates the analytical relationship between the integrated quality and the number of tables, for different individual table qualities.

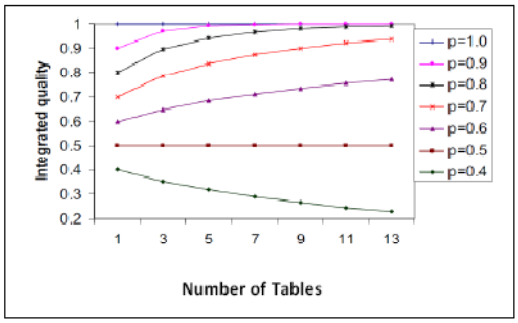


Fig. 1: The relationship between integrated quality, individual quality, and the number of Tables.

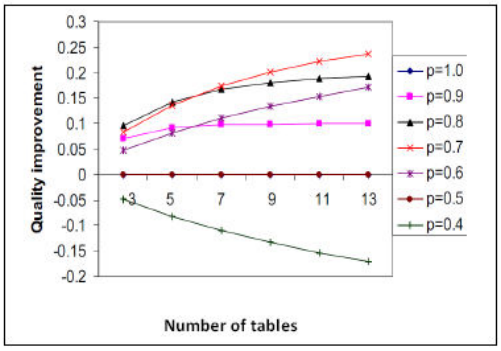


Fig. 2 : Improvement in integrated quality compared to single-table, as a function of the number of tables, for different tables.

As expected, the integrated quality improves with larger numbers of tables, when the individual table quality $p > 0.5$; however, the marginal improvement decreases as the number of table increases. Moreover, the benefit of getting more tables also depends on the underlying value of p. Figure 2 shows how integrated quality q increases compared to the case of single-labeling, for different values of p and for different numbers of tables. For example, when $p = 0.9$, there is little benefit when the number of tables increase from 3 to 11. However, when $p = 0.7$, going just from single labeling to three tables increases integrated quality by about 0.1,

(b). *Different Tables Quality*

If we relax the assumption that $pj = p$ for all j, and allow tables to have different qualities, a new question arises: what is preferable: using multiple tables or using the best individual table? A full analysis is beyond the scope (and space limit), but let us consider the special case that we have a group of three tables, where the middle labeling quality is p, the lowest one is $p - d$, and the highest one is $p + d$. In this case, the integrated quality q is:

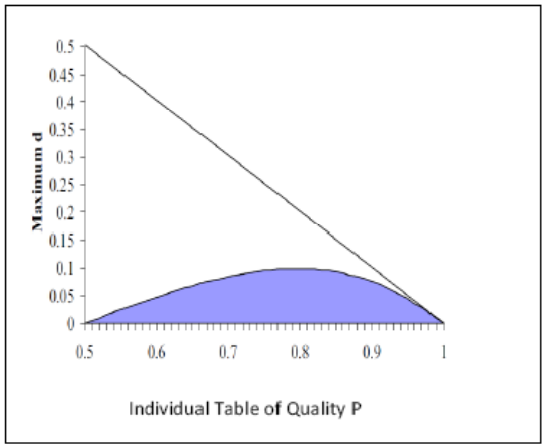


Fig. 3 : Repeated-tables improves quality when d is below the curve

When is this quantity greater than that of the best table $p + d$? We omit the derivation for brevity, but Figure 3 plots the values of d that satisfy this relationship. If d is below the curve, using multiple tables improves quality; otherwise, it is preferable to use the single highest-quality table.

(c). *Improvement Data Quality tables*

The example above suggests a straight forward procedure for selective repeated-labeling: acquire additional tables for those examples where the current multi set of Data is improved.

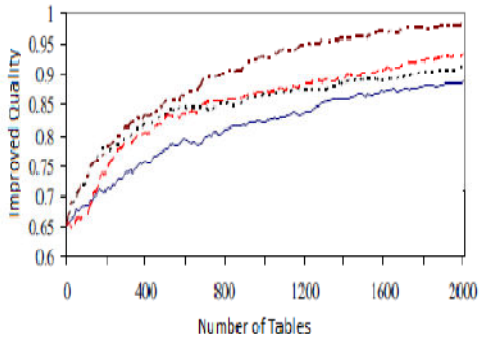


Fig. 4 : The data quality improvement of Tables

VII. CONCLUSIONS

Considering the progressing technology and facilities, and having a huge amount of data in the shortest time, there is a need to have some facilities to analyze these data fast and accurately. On the other hand, the mistakes while recording data, or selling products despite technology is inevitable. Therefore, we need to decrease mistakes in data extraction of relational tables and have more control on the products of companies in order to have reliable information and data. Since most of the databases in factories and industries are relational and these relational tables are related, MRDM can help to diagnose these mistakes. The future work improves more and more reliable using association rules in Multi-relational Data Mining for any production company.

REFERENCES

[1] Arno Jan Knobbe "Multi-Relational Data Mining" Faculteit Wiskunde en Informatica, Universiteit Utrecht, pp. 10, Nov. 2004.

[2] Mary DeRosa'' Data Analysis for Counterterrorism '' Library of Congress Cataloging-in-Publication Data CIP information available on request, pp. 1-33, March 2004.

[3] Miao Liu, Hai-Feng Guo, Zhengxin Chen" On Multi-Relational Data Mining for Foundation of Data Mining'' IEEE International Conference, pp. 389-395, 2007.

[4] Dasu, T., Johnson, T. Exploratory "Data Mining and Data Cleaning, Wiley, 2003.

[5] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Advances in Knowledge Discovery and Data Mining, MIT Press, 1996.

[6] Hand, D., Mannila H., Smyth, P., Principles of Data Mining, MIT Press, 2001.

[7] Holsheimer, M., Kersten, M., Mannila, H., Toivonen, H. A Perspective on Databases and Data Mining, In Proceedings of KDD '95, 1995

[8] Guoling Liu, Runian Geng" An Efficient Algorithm for Mining Association Rules from Multiple Databases" 2nd International Conference on Computer Engineering and Technology (ICCET), vol. 4, pp. 349-351, April 2010.

[9] Tamrapani Dasu and Theodore Johnson "Exploratory Data Mining and Data Cleaning" Journal of Statistical Software, vol. 11, September 2004.

[10] Date, C. An Introduction to Database Systems, Volume I, the Systems programming Series, Addison-Wesley, 1986.

[11] Ullman, J., Widom, J., A First Course in Database Systems, Prentice Hall, 2001.

[12] Eric Ka Ka Ng, Ada Wai-Chee Fu, KeWang''Mining Association Rules from Stars'' Proceedins IEEE International Conference, pp. 322-329, 2002.

[13] Amanda Clare, Hugh E. Williams Nicholas Lester'' Scalable Multi-Relational Association Mining'' Proceedings of the Fourth IEEE International Conference on Data Mining, IEEE Computer Society, pp.355-358, 2004.

[14] Jingfeng Guo, Weifeng Bian, Jing Li" Multi-relational Association .Rule Mining with Guidance of User" Fourth International Conference on Fuzzy Systems and knowledge Discovery, IEEE Computer Society ,vol. 2, pp. 704-709, 2007.

[15] Hong Li, Xuegang Hu" Efficient Mining of Strong Negative Association Rules in Multi-Database" Computational Intelligence and Software Engineering, pp. 1-4, Dec. 2009 .

[16] Yun Li, Luan Luan, Yan Sheng, Yunhao Yuan" Multi-relational Classification Based on the Contribution of Tables" International Conference on Artificial Intelligence and Computational Intelligence, IEEE Computer Society, vol. 4, pp. 370-374, Nov. 2009.

[17] Guoling Liu, Runian Geng" An Efficient Algorithm for Mining Association Rules from Multiple Databases" 2nd International Conference on Computer Engineering and Technology (ICCET), vol. 4, pp. 349-351, April 2010.

[18] Mahmoud Houshmand, Mohammad Alishahi "Improve the classification and Sales management of products using multi-relational data mining" IEEE 3rd International Conference on Communication Software and Networks (ICCSN) , pp. 329-337, May 2011.

[19] Saeed Farzi, Ahmad Baraani Dastjerdi" Data Quality Measurement using Data Mining", International Journal of Computer Theory and Engineering, Vol. 2, No. 1 February, 2010.

Issues of Emotion-Based Multi-Agent System

Arushi Thakur[1] & Divya Rishi Sahu[2]

[1&2]CSE - Department, IES College, Bhopal

[1]IES Group of Institute, Bhopal - 462052 (M. P.), India, [2]S-671, Nehru Nagar, Bhopal - 462052 (M. P.), India

Abstract - Emotion plays a significant contribution in perceptual processes of psychology and neuroscience research. Gently, area of Artificial Intelligent and Artificial Life in simulation and cognitive processes modeling uses this knowledge of emotions. Psychology and neuroscience researches are increasingly show how emotion plays an important role in cognitive processes. Gradually, this knowledge is being used in Artificial Intelligent and Artificial Life areas in simulation and cognitive processes modeling.

Researchers are still not very clear about working of mind to generate emotion. Different people have different emotion at the same time and for same situation. Thus, to generate artificial emotion for agents is very complex task. Each agent and its emotion are autonomous but when we work on multi-agent system. Agents have to cooperate and coordinate with each other.

In this paper we are discussing the role of emotions in multi-agent system while decision making, coordinate and cooperate with other agents. Also, we are about to discuss some major issues related to Artificial Emotion (AE) that should be considered when any research is proposed for it.

In this paper we are discussing the role of emotions in multi-agent system while decision making, coordinate and cooperate with other agents. Also, we are about to discuss some major issues related to it that should be considered when any research is proposed for Artificial Emotion (AE).

Keywords - *Emotion, Artificial Intelligence, Multi-agent System, Artificial Emotions.*

I. INTRODUCTION

Areas of computer science such as Artificial Intelligence and multi-agent system employed heuristics like human emotions to organize their complexity. Human emotions play a vital role in area of artificial intelligence as heuristics, however human try to resolve problem. 'Marvin Minsky boldly' stated that "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions" [1]. Emotions are a crucial part of the believability of characters that interact with humans [6, 3]. Multi-agent systems provide interface where individual agent is supposed to be autonomous. Applications related to practical and real life based on decision-making concepts (probability and utility) and rules (maximizing expected utility) will leave an agent with multiple actions or plans with slightly equal preference. This gives rise to no determinism in an agent's decision making, which is a problem if the agent has to choose only one of the options. In this nondeterministic state each agent has to interact with other agents. Most of the time human involves emotion in decision making as well as interaction and communication with others. Human in communication process utilize messages that contain four factors: facts, relationship, appeal and self-revelation [5]. Emotion is a crucial element to model perception, learning, decision process, memory, behavior and others functions they are interested in.

II. ARTIFICIAL EMOTION

Different explanations of emotion are proposed from different point of views as neuroscience, philosophy, social and culture studies. Different theories focus on different aspects of emotion. But in broader sense they all share a common concept that is known as FUNCTIONAL VIEW of emotion. It says that emotion is an output (or action) to serve a purpose or to satisfy its environment for a given set of input (or events).

When any event occurs, emotion plays a role to react. There are two ways of how to react. First way is to think all the output first and then conclude which one easy to react or implement. On the other hand second way is to compare the event with the previous event's database and then decide the action. In both of the cases system has a purpose to serve and an environment to satisfy and to act with.

Role of emotion in biological point of view is still not very clear. Different people react different in same situation. There are three layers of behavior. First is memory of previous events, Second layer is mood and third layer is personality. Thus, Emotions differ from person to person at the same time and same condition. So when we talk about artificial emotion, to construct a

set of subsystems which will produce emotion, is very complex.

III. COOPERATION & COORDINATION AMONG AGENTS

Since a long time Human Computer Interaction (HCI) community is trying to make a system that can interact with others with emotions as human do. Now a days, Some robots have been introduced who communicate with users and its results are also very satisfactory. Some review articles for this are also available [4,7]. Now, in multi agent system, If agents, which have to cooperate with each other, know how they work internally, they can anticipate expected actions. For example: If an agent know that another agent working on a task to achieve a common goal and also having so much work load, at this time, an agent can contribute or coordinate with the other agent to avoid failure. An agent could achieve this by mapping the perceived state of another agent to its own mechanism and approximate the affective state of the that agent and by this mechanism it can then predict what actions the other agent is most likely to perform next after decision making by using of heuristic. If all agents were using an affective mechanism as a heuristic for their decision making, this kind of cooperation can be achieved in much the same way as humans do in their interactions.

IV. MAJOR ISSUES

For some years, experimental research using emotion-based agents is being developed. We could mention [2], which measure emotion as behavior modulation. Next [8] in which, different levels of an artificial hormone mechanism generate emotion.

A. COMPLEX TO UNDERSTAND

Considering current state of the running projects, to make an emotion-based system is far from simple and straightforward job. Even computational concepts of emotion are as problematic and complex as computational understandings of human being's emotion and their normal life.

B. LACK OF FRAMEWORK

A well defined scientific framework is needed for achieving "Artificial Emotions". Some research works (e.g. [12], [10], [9]) show advanced knowledge and concepts to follow, approaches that might be successfully used to model artificial emotions in agents.

C. LACK OF TRUST

Another issue is lack of trustworthy result. The question is "how much can we trust on an emotional decision?" When an agent predict the action that another agent is most likely going to perform, and take an action for coordination, so how much is it trustworthy that the another agent would really take the same action that is predicted.

D. COMPARISION STUDY

Another important issue is to comparisons between projects and also within same project, with comparative results from emotion and non-emotion-based experiments.

V. ISSUES WITH ARTIFICIAL EMOTION

When we talk about artificial emotion, we have to decide the levels and limit of emotion. For example if we say "care" emotion so at what extent system should express care emotion. In real life, if a person is too much caring for someone, it becomes possessiveness and too much possessiveness becomes hard to accept and sometime dangerous also. So caring should have some limit. In the same way there should be some limit in artificial emotions. But the problem is how you would decide the levels. Or how will you calculate the levels. What would be the "primitives" and "starting point".

Some authors say that there are two important things: set of inputs (events) and environment. Now design architecture to serve a purpose and act as output. It would have some primary emotional responses and then leave the subsystems to cooperate and produce a final emotion and action. Some can argue that there should be a mechanics in each agent's subsystems that will produce emotions using primary emotional responses. But problem with this approach is Uncertainty. How would you know what the system will react after an event. We will never know whether we will get what we want or not.

Another major issue to test the correctness of emotion, it can be done by comparing the outcomes from the system behavior with the outcomes that would be produced by equivalent biological system. But it is very hard to implement.

VI. FUTURE RESEARCH AND CONCLUSION

We are currently working on the OCC model that is trying to formalize and implement 22 deliberative emotions [10]. OCC model is used as heuristics for controlling the situation of no determinism in goal-directed agents. The OCC model has been used for emotion synthesis. This model is very suitable for formalization but we focus to broaden our research with other alternative theories of emotion. Recently, we are formalizing the whole set of 22 emotions in an extension of the KARO framework [13, 11]. We expect to accomplish this by allowing agents to map the perceived states of other agents to their own

affective model, which will allow agents to approximate the goals and plans of other agents as they reason about their own. As soon as neuroscience researches increase, they might be more and more useful in the construction of emotion-based multi-agent systems. Computational projects with particular focus will be able to extend their scope to achieve this goal.

In this paper, we have mentioned three problems in emotion-based multi-agent systems, namely no determinism situation in decision making, cooperation and coordination in multi-agent systems, and believability.

By overcome these challenges can be a very important step to go beyond currently available engineering applications and towards a more scientific discipline of computer science ([14]).

VII. ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

REFERENCES

[1] Minsky, M.L.: The society of mind. Simon and Schuster, New York, N.Y. (1986)

[2] Kato, Tomoko and Arita, Takaya. (2005) "Evolutionary Simulations based on a Robotic Approach to Emotion". The 10th International Symposium on Artificial Life and Robotics. Oita, Japan, February, 2005.

[3] O'Reilly, W.S.N.: Believable Social and Emotional Agents. Carnegie Mellon University, Pittsburgh, PA (1996)

[4] Bartneck, C., Okada, M.: Robotic User Interfaces. Human and Computer Conference (HC2001), Aizu (2001) 130-140

[5] Schulz, F.v.T.: Miteinander Reden - Stoerungen und Klaerungen. Rowolth Taschenbuch Verlag GmbH, Reinbeck bei Hamburg (1981)

[6] Elliott, C.D.: The Affective Reasoner: A Process model of emotions in a multi-agent system. The Institute for the Learning Sciences, Northwestern University, Evanston, Illinois (1992)

[7] Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems 42 (2003) 143-166.

[8] Gomi, T. and Ulvr, J. (1993) "Artificial Emotions as Emergent Phenomena". In Proceedings of IEEE Workshop on Robot and Human Communication, Tokyo, Japan, November, 1993.

[9] Arbib, M. A. and Fellous, J-M. (2004) "Emotions: from brain to robot". Trends in Cognitive Sciences. V. 8 no 12, pp. 554–561

[10] Sloman, A. (2002) "How many separately evolved emotional beasties live within us?" In Trappl, R., Petta, P. and Payr, S (eds.) Emotions in Humans and Artifacts. The MIT Press, pp. 35–114

[11] J.-J. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. Arificial Intelligence, 113:1–40, 1999.

[12] Cañamero, D. (1997) "Modeling Motivations and Emotions as a Basis for Intelligent Behavior". In: First Conference on Autonomous Agents, 2001, Marina Del Rey, California: ACM. 1997. pp. 148-155.

[13] John-Jules Ch. Meyer. Reasoning ´about emotional agents. In Ramon L´pez o de M´ntaras and Lorenza Saitta, editors, Proceedings of the 16th Eureopean Conaference on Artificial Intelligence (ECAI 2004), pages 129–133. IOS Press, 2004.

[14] AAAI press (2004). In Hudlicka, E. and Cañamero, D. (eds.) "Architectures for modeling emotion: cross-disciplinary foundations". 2004 AAAI Spring Symposium Technical Report Available at :

◈◈◈

Feature Extraction Based Face Recognition Using Extreme Learning Machine (ELM)

Nagabhairava Venkata Siddartha, Mohammad Umar, Nabankur Sen & P.Krishna Prasad

CSE, SRM UNIVERSITY

Abstract - In recent years, Face recognition becomes one of the popular biometric identification systems used in identifying or verifying individuals and matching it against library of known faces. Biometric identification is an actively growing area of research and used in electronic commerce, electronic banking, electronic passports, electronic licences and security applications. Face recognition finds its application in wide variety of areas like criminal identification, human - computer interaction, security systems, credit- card verification, teleconference, image and film processing. This paper suggests an automated face recognition system which extracts the features from the face. Feature extraction process includes locating the position of eyes, nostrils and mouth and determining the distances between those regions. From the extracted features, a database is created for known individuals. A virtual neural network is created based on Extreme Learning Machine (ELM).

Keywords - *Biometric identification, Face recognition, Feature extraction, ELM, FAR, FRR.*

I. INTRODUCTION

Biometrics is a measurable physiological or behavioral characteristic of an individual used in personal identification and verification. It includes fingerprint, iris, face, voice, palm symmetry, hand geometry and so on. Biometric identification has significant advantages over other authentication techniques because biometrics characteristics are not easily modifiable and are unique. Fingerprint recognition has been widely used because it is cost affordable and best utilized in small-scale verification systems. This recognition method finds applications in mobile phones, computers, Employees identification scheme, etc. But this method encounters problems like some fingerprints are unsuitable for use due to cuts or other defects. Also artificial finger straps are readily available in the market makes the recognition process difficult or identifying the wrong individual. Iris recognition has evolved in recent years which eliminate the problem in fingerprint mechanism. The accuracy and speed of iris systems allows this technique implementing in a large scale system. The iris of each person is distinctive and even identical twins have different patterns. Since it is extremely difficult to alter the texture of the iris through surgery, it would be difficult for someone to provide wrong identifications. Also it is relatively easy for the system to detect when an artificial iris specially made by contact lens, is being used to gain identification. But iris recognition also encounters some difficulties in the verification applications. To overcome such difficulties in fingerprint and iris recognition techniques, Face recognition comes into existence in the modern world of artificial intelligent systems.

1.1 FACE RECOGNITION

Face recognition field has achieved a significant growth over the past few years. It is the popular area of research for more than 3 decades in computer vision and the most successful applications of image analysis. Several companies offer face recognition software that can produce high-accuracy results with a large database. Recent research involves developing approaches that accounts for changes in lighting, expression, and aging, for a given person. Also, researches under this field include dealing with glasses, facial hair, and makeup.

Two predominant approaches in face recognition system are geometric feature- based and appearance-based. The geometric feature based approach uses the properties of facial features such as location of eyes, nose, mouth, chin and their relations for face recognition descriptors. The appearance-based face recognition approach operates directly on image based representation. The whole face region is the raw input to a recognition system and the facial features are processed as templates. Face recognition is commonly used in two ways, Face identification and Face verification [8]. First refers one to many matches and next refers one to one matching. The automated methods of facial recognition work very well, but it do not recognize persons

effectively as a human brain. Regarding face recognition problems, it also encounters the combined variations in illumination, pose, expression, spectacles, and optimization of training databases and also needs the real-time requirements.

1.2 EXTREME LEARNING MACHINE (ELM)

ELM is one of the virtual neural networks, provides less training time and high accuracy. It is a sequential learning algorithm where the training observations are sequentially used as single data block or data with varying or fixed length in the learning algorithm. At any time, only the new observations are seen and learned. The training data are discarded as soon as the learning procedure for that particular data is completed [4].

In this project, ELM is used for training and testing databases of the face images. For recognition process, 35 images are taken into consideration. 9 images are used for training process and the remaining for testing process. Number of hidden nodes is manually entered. Input weights and bias are randomly assigned based on the inputs and hidden neurons. The created face database is trained using ELM network and matching is performed with test images. ELM works well even for small set of database. As the number of hidden neurons gets increased, high accuracy is achieved.

II. LITERATURE SURVEY

Atefe Assadi and Alireza Behrad [1] proposed a method for Face recognition using Texture and depth information. This method provided a 3D approach for recognizing faces under pose variation and different illumination conditions. Scale-invariant feature transform (SIFT) descriptors are used to extract the facial feature points and compared with the database. They also calculated the matching points using SIFT feature vector. Input face image with maximum matching points is recognized as known face.. This method provided 88.96% recognition rate.

Ramesha K et al [5] proposed a Feature Extraction based Face Recognition, Gender and Age Classification (FEBFRGAC) algorithm. In this paper, recognition process was performed based on the geometric features based on the symmetry of human faces and the variation of gray levels, the positions of eyes, nose and mouth were extracted and located by applying the Canny edge operator. The gender was classified based on posteriori class probability and age was classified based on the shape and texture information using Artificial Neural Network. This algorithm provided face matching ratio is 100%, gender classification is 95%, and age classification is 90%.

Shermina J [6.a] proposed a face recognition system based on Multi linear principal component analysis (MPCA) and Locality preservation projection (LPP). In this paper, after face image preprocessing, dimensionality reduction is performed using MPCA. Features were extracted using LPP which provided nearest neighbor search in the low dimensional space. Recognition was performed by using L2 similarity distance measure, computed between the database image and the query image. High recognition rate was achieved by combining both the MPCA and LPP.

Shermina J [6.b] proposed a Face recognition system based on Discrete Cosine Transform (DCT) and PCA. In this research paper, low frequency DCT components are used to normalize the illuminated image. 64 illumination conditions are taken into account. This paper provided with accuracy of 94.2% and concluded that combination of DCT with any other recognition methods provided significant illumination invariant recognition accuracy.

Thamizharasi A [8] proposed a survey paper of Analysis on Face recognition by combining multi scale techniques and Homomorphic filter using Fuzzy k nearest neighbor classifier. In this paper, DCT and Discrete wavelet transform (DWT) were the two multi scale techniques used. Homomorphic filters were used for normalization of illumination. K means clustering algorithm was applied to group the pixels in the preprocessed image based on gray-scale threshold values. Fuzzy k nearest neighbor classifier was used to classify image in the test database by calculating the Euclidean distance matrix within the train database. 2D Haar DWT at level 1 was performed on the preprocessed image for choosing the approximate coefficients at level 1. Then the clustering algorithm and classifier were used for finding the face recognition rate. High recognition rate was achieved by combining all these multiscale techniques even though they could be used individually. DCT yielded 89.5% recognition rate while DWT yielded 90% rate with Homomorphic filter, K means clustering and Fuzzy k nearest neighbor classifier. The system became more complex because of more no. of techniques and computation time would be more.

III. PROPOSED METHOD

This paper proposed a method for feature extraction based face recognition using ELM network. The recognition process will be used for end user security in image authentication systems. Medical images are confidentially transmitted via wireless channel with higher level of security using various types of encryption/decryption algorithms. These algorithms avoid hacking of medical images while transmission across internet. But there is no security at the end user or receiver's side and thus anyone can receive the

encrypted message. The encrypted key will be public means the intruder decrypts the message and gets the medical image. Hence it is necessary to provide authentication, face recognition algorithm is provided at the end users side. The flowchart describes the proposed model will be shown in the figure1.

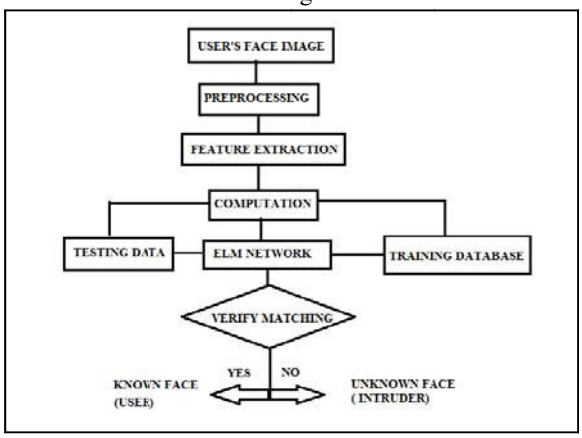


Fig. 1 : Face recognition model

Initially an RGB colour image of the user is captured using a web camera or high resolution video camera. The size of the face image is 640x480. It is represented as an array of mxnx3 color pixels corresponding to the red, green and blue components of an RGB image. Three dimensional RGB is converted into two dimensional binary image based on threshold values .The input face image is transformed into binary face image for retaining the important features. Background changes, illuminations are adjusted and concentrating on the face region alone. This process referred as Preprocessing for improving the quality of the image shown in the figure 2.

Fig. 2 : RGB image of user

Feature extraction is performed after preprocessing the input face image. The regions of two eyes, nostrils and mouth are located in the face image and Blob measurement properties are used in estimating the connected components in the face image. Thus the eyes, nostrils and mouth regions are located in the binary image. Then the regions are represented within the bounding boxes and filled the disconnected pixels are shown in the figure 3.

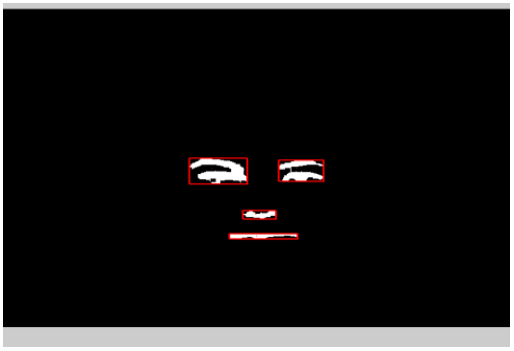


Fig. 3 : Detected Regions with bounding boxes

From the extracted regions, the following distances are measured in the face image.

Inter-Ocular Distance - The distance between the right eye and the left eye pixels.

Eye to Nose Distance - The distance between the midpoints of the line joining the eyes and the nose tip pixels.

Eye to Mouth Distance - The distance between the midpoint of the line joining the eyes and the center point of the mouth.

Nose to Mouth Distance - The distance between the nose tip and the center point of the mouth.

The above said distances are calculated from the face image and the ratios are calculated. These computed ratios are referred as features of the face image which are taken into account for recognition.

The ratios are mentioned as follows,

1. **EENR** is the ratio between the Inter-ocular distance and the Eye to Nose distance.
2. **EEMR** is the ratio between the Inter-ocular distance and the Eye to mouth distance.
3. **EENMR** is the ratio between the Inter-ocular distance and Nose to mouth distance.
4. **ENEMR** is the ratio between the Eye to Nose distance and Eye to mouth distance.

Face image	EENR	EEMR	EENMR	ENEMR
06-1m	1.1895	0.6548	1.4571	0.5505

11-1m	1.1602	0.8944	3.9055	0.7709
14-1f	1.4506	0.8730	2.1926	0.6018
15-1f	1.3810	0.9320	2.8666	0.6748
20-1m	2.8710	1.7429	4.4360	0.6070
21-1m	1.1133	0.6869	1.7936	0.6170
25-1m	1.4453	0.7610	1.6074	0.5265
28-1m	1.2863	0.8328	2.3626	0.6474
29-1m	1.3404	0.8107	2.0809	0.6048

Table 1. Database for different face images

The table 1 shows that the extracted features of the sample database images used in the recognition process.

The computed features from the face image are given as inputs to an extreme learning machine network. Number of hidden layer neurons alone is entered manually or it will be the sum of input neurons and output neurons. Input weights and biases are assigned randomly and from that output weights are calculated. The features are trained within the network for the given database. The query image is verified for matching purpose. If matching exists, the result shown as *KNOWN FACE* otherwise the result will be *UNKNOWN FACE*. Thus User verification is performed once the extracted features are matched with the database otherwise user cannot access the authority to use the resources. Figure 4 shows the simple architecture of ELM network.

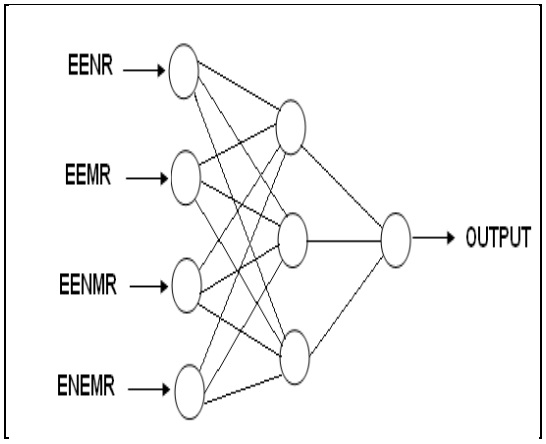


Fig. 4 : A simple ELM network

In this proposed method, 35 face images of different users are used for recognition purpose. Out of these 35 images, 9 images are taken as training data and remaining 26 images are taken for testing data. The accuracy or performance metrics of the biometric identification depends on two parameters i.e. FAR and FRR.

False Acception Rate or False Match Rate (FAR or FMR) is the probability that the system incorrectly matches the input pattern to a non-matching template in the database. It measures the percent of invalid inputs which are incorrectly accepted. False Reject Rate or False Non-Match Rate (FRR or FNMR) is the probability that the system fails to detect a match between the input pattern and a matching template in the database. It measures the percent of valid inputs which are incorrectly rejected.

IV. RESULTS AND CONCLUSION

The proposed face recognition method yields the best authentication system. This method is simple and efficient. It deals with pixel information rather than texture information hence the accuracy will be more. The use of ELM network for training and testing the database provides fast and accurate authentication system. The measured FAR is 3.85% and FRR is 0% hence the proposed face recognition system yields a better biometric identification system using ELM network

REFERENCES

[1] Atefe Assadi and Alireza Behrad "A new method for human face recognition using texture and depth information", IEEE transactions on Neural Network Applications in Electrical Engineering (NEUREL), pp. 201-205, 2010.

[2] Deo Brat Ojha, Ajay Sharma, Abhishek Dwivedi, Nitin Pandey, Amit Kumar, "An Authenticated Transmission of Medical Image with Codebase Cryptosystem over Noisy Channel", International Journal on Advanced Networking and Applications ,vol. 02, Issue: 05, pp. 841-845,2011.

[3] Deo Brat Ojha et al , "An Authenticated two-tier security on transmission of medical image using codebase cryptosystem over teeming channel", International Journal of Computer applications, vol.12-no.9, pp. 22-26,2011.

[4] Nan-Ying Liang, Guang-Bin Huang, P.Saratchandran, and N. Sundararajan , "A fast and accurate online sequential learning algorithm for feed forward networks", IEEE transactions on neural networks, vol. 17, no. 6,2006.

[5] Ramesha K et al, **"**Feature Extraction based Face Recognition, Gender and Age Classification"**,** International Journal on Computer Science and Engineering (IJCSE), vol. 02, no.01S, pp.14-23, 2010.

[6.a] Shermina J, "Face recognition system using multilinear principal component analysis and locality preservation projection", IEEE GCC conference and exhibition, Dubai, United Arab Emirates, 2011.

[6.b] Shermina J , "Illumination invariant face recognition system based on Discrete Cosine Transform and Principal Component Analysis", International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), pp. 826-830, 2011.

[7] Sunanda Mulik, "Face recognition for biometric identification: A review", International Journal of Intelligent Information Processing, vol.4, pp. 89-92, 2010.

◈◈◈

Complete Automation of Metro Stations through Artificial Intelligence

Rittick Datta & Prachi Taksali

Computer Science Engineering, University of Petroleum and Energy Studies (UPES), Dehradun, India

Abstract - Metro stations have become an invaluable transportation resource and will be spreading out of the metropolitan cities soon. It has reduced travel time and travel cost. We intend to research the possibility of unmanned metro stations through the application of artificial intelligence, one of which is expert systems. Expert systems –that are able to hold the accumulated knowledge of different domain experts can be implemented to guide the commuter about the optimum travel route. In this way the metro stations can be turned into self-sustainable structures.

Keywords - *metro stations, expert systems, artificial intelligence.*

I. INTRODUCTION

A metro station is an electric passenger railway for rapid transportation in urban areas. It comprises of underground tunnels, elevated rails and is multi-level at the stations. The metro station has several entrances for ease of access. The commuters are directed towards the entrances with the help of logo marks. It is connected to significant buildings by direct enclosed hallways .The metro stations also exhibit art and beautiful architecture. It has reduced travelling time and cost for the public. Its contribution to curb pollution is significant when the issue of global warming is alarming. In some metro stations the entire platform is screened from the tracks by a glass wall and consists of automatic platform-edge doors. In such a scenario, it becomes mandatory for the approaching train to arrive at a slow speed to halt in alignment with the doors. Ventilation of the platform is taken care of according to the weather, heated or cooled. Metro railway has high capacity and frequency to cater to the public. It has grade separation from other traffic. Grade separation refers to the method of aligning a junction of two or more surface transport axes at different heights.

Fig. 1: Underground tunnel

Now, we will consider an example of the Delhi Metro to analyze the figures associated with a metro station. Later, we will talk about, why we should automate and then discuss how we can automate.

Fig. 2: Elevated rails

II. CASE STUDY: The Delhi Metro

The Delhi Metro serves Delhi, Gurgaon, Noida and Ghaziabad. The network consists of six lines. It has a total length of 189.63 kilometers. There are 142 stations out of which 35 are underground. It is a combination of elevated, at-grade (on the same level) and underground lines. The Delhi Metro was built and is operated by the Delhi Metro Rail Corporation Limited (DMRC). DMRC operates around 27000 trips between 06:00 hrs – 23:00 hrs with an interval of 2 minutes 30 seconds between trains at peak frequency and otherwise with an interval between 3 minutes – 4 minutes 30 seconds.

Fig. 3: Platform screened from the tracks

Due to population increase in the city, number of coaches has been increased from 4 to 6. The four lines are Red line(Dilshad Garden to Rithala),Yellow line (Jahangirpuri to Huda city centre),Blue line (Dwarka sec-21 to vaishali/Noida City Centre) and Violet line (Central Secretariat to Badarpur).Power output is supplied by 25 Kilovolt, 50 Hertz alternating current through overhead catenary which are overhead wires used to transmit electrical energy .

Fig. 4: Multi-level at stations

On a daily basis, Delhi Metro has 1.6 million commuters. It is also certified by United Nations as the first metro rail and rail based system in the world to get "Carbon credits for reducing greenhouse gas emissions" and helping in reducing pollution levels in the city by 6.3 lakh tonne every year.

Fig. 5: Art exhibited at Mandi House metro station

Metro stations have services like ATM, food outlets, cafes and convenience stores. Eating, drinking, smoking and chewing of gum are strictly prohibited in the entire system to maintain order among the commuters. Sophisticated fire-alarm system is available for advance warning in case of emergency.

Over 3500 CISF personnel have been deployed to deal with law and order issues in the system, in addition to metal detectors, X-ray baggage inspection systems and dog squads. Intercoms are provided for emergency communication between passengers and Train operator. Periodic security drills are carried out at stations and on trains.

For ticket purchase, passengers can opt for either RFID (Radio Frequency Identification) token or Smart card. Tourist cards are also available.

Fig. 6 : RFID token and Smart card

III. RFID (Radio Frequency Identification) token

It uses a wireless non-contact radio system. Two-way radio transmitter-receivers called readers send a signal to a tag and read its response .The readers transfer the gathered information to computer systems running the RFID software. The tag's information is stored electronically in a non-volatile memory. The tag includes a small RF transmitter and receiver. An RFID reader transmits an encoded radio signal to interrogate the tag. The tag receives the message and responds with its identification information. The tag need not be in sight of the reader but can be embedded and can be read from several meters.

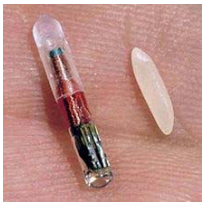


Fig. 7: RFID chip

IV. WHY COMPLETELY AUTOMATED METRO STATIONS?

- Machines can perform monotonous and tedious tasks repeatedly with tireless precision .It will completely remove any room for error as in case of human beings.

- It will increase the efficiency and effectiveness of the system as a whole.
- Travelling time can become more precise as the system will move towards complete automation without any human interference.
- Since work of an employee at a ticket counter is repetitive, it can be replaced by machines and the skills of that human resource can be invested elsewhere where it is more required. It will also cut-down expenses of the corporation investing in metro stations, since employees working at metro stations will converge to none.
- This can also bring down the ticket prices and make metro a more favourable mode of transportation among the public which will lead to lesser automobiles on the road, hence a greener planet.
- The installation of solar panels in place of glass windows can contribute in making it a self sustainable structure by lowering electricity bills and relying on power supply only for trains but not for lighting up the metro station approximately 17 hrs on a daily basis.
- The (self sustainable structures) metro stations will keep each and every other metro station informed about the arrival and departure of the trains, like human entities but in precise real-time and help curb power consumption further.
- If the security can also be automated using high end security devices and restricting Central Industrial Security Force (CISF) to patrolling and interfering only in case of an emergency, their skills can be put to better use in more violent parts of the country.
- It will be a onetime high investment with the vision of an economical long run.

V. MANNED LOCATIONS IN A METRO STATION TO BE AUTOMATED

- Ticket counters
- Security
- Shops

A. Use of KBS at ticket counters

The key idea is to separate the knowledge about the problem from the process by which the problem is solved. The separation of knowledge from control makes it easier to add new knowledge or remove existing knowledge if necessary. These systems have three basic components: a knowledge base in the form of facts, rules, frames or objects; an inference engine in the form of algorithms on how to control the processing; and a database.

The problem to be handled at a ticket counter is to suggest the shortest possible path from source to destination to the passenger. All the permutations and combinations of the routes connecting each and every metro station to the other on the network are to be stored in the form of rules in the Knowledge Base System.

The commuter will approach the ticket counter and interact with the Knowledge Base System. She/he will specify the source and destination .The inference engine will process the request and search through the Knowledge Base System for the best possible route in terms of time and travelling cost. The user can choose an intermediate stop in the above request which will act as a constraint and will have to be satisfied for the result to be valid.

If the time and cost factors of two or more routes are close to each other and the system is in ambiguity, it can display the short listed optimum possible routes for the user to choose one.

Once the passenger has enquired about the shortest path from source to destination, she/he will interact with the automatic TVM (Token Vending Machine).The Token Vending Machine will be the process of solving the problem. It will be coin-operated/Smart card operated and will dispense valid token after the commuter has entered the source and destination and has confirmed selections. After receiving the token, the person can move towards the platforms after verifying the token at the flap gate.

B. Security

The key responsibility of security is to detect passengers carrying arms and ammunition, hence removing the possibility of threat. They use metal detectors and X-ray baggage inspection systems. The way to automate security with least amount of human intervention is as follows:

Interchange the sequence of purchasing of token and passing through security check, i.e. every person entering the metro station will have to clear a security check which will be a very speedy process. Please see the diagram below in order to understand the process.

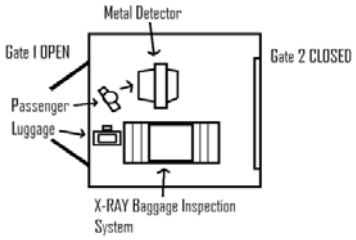


Fig. 8: Passenger enters the security check area.

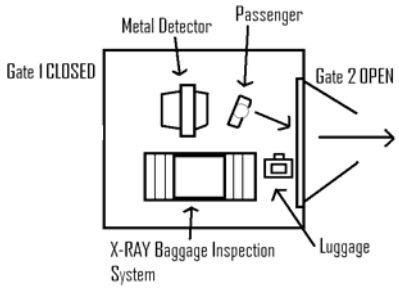


Fig. 9 : Passenger clears the security check and moves towards automated ticket counter

The Gate 1 (entrance to the metro station) will be open and will allow only one passenger to enter the security check area .The passenger will be monitored by close-circuit cameras for suspicious behavior in case of which the security patrol will be informed via telecommunication with a recorded message. The passenger will place his luggage on the X-Ray baggage inspection system and will then move past the metal detector to clear the security check. Only after the two checks are cleared the Gate 2 will open and the passenger can move towards the automated ticket counter. Please note that while the person is undergoing security check gate 1 is closed until that person clears the security check.

In case a threat has been detected, the two gates of the security check area will be locked automatically trapping the culprit. This can be implemented with the help of sensors and actuators installed inside the security check area. An alarm will go off; warning the public of the danger and a pre-recorded SOS message will be sent to all security units in and around the metro stations.

C. Automated Vending Machines

Food outlets and cafes can be replaced by AVMs (Automatic Vending Machine) .Although it isn't much of an application of artificial intelligence but it serves the purpose of making metro stations unmanned to an extent, hence appropriate. These vending machines are coin-operated or ATM card operated.

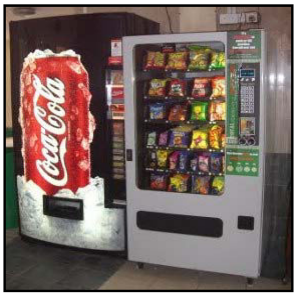


Fig. 10 : Automatic Vending Machines instead of food outlets

D. Solar panels to save electricity

For the metro station to be completely self sustainable after the ticket counter , security and shops have been automated to a great extent , it is required that they produce their own electricity for purposes other than running trains and be least dependent on power lines.

One way to achieve this is by replacing the glass panes by solar panels. The photo voltaic cells of the solar panels will directly convert sunlight into electricity and save power. It is an important step towards self sustainability because all other systems which will be responsible for automating metro stations run on electricity.

Fig. 11: Solar panels can replace glass window panes.

VI. CONCLUSION

Metro stations are restricted to metropolitan cities for now but it is likely that in the near future, each urban area might have a metro station for public transportation. By utilizing artificial intelligence and automating metro stations, we can make better use of human resource and minimize pollution. The different metro stations will coordinate among themselves through telecommunication more precisely in real time than human beings do with time lag. The frequency of trains will be a function of the rush observed at different times of the day avoiding unnecessary schedules and cutting down power consumption.

It will be a onetime high investment aimed at an economical long run.

VII.ACKNOWLEDGMENT

The authors would like to thank Dr. Neelu Ahuja for the sound concepts developed in the course of artificial intelligence because of which we were able to think of its application in metro stations. We would also like to thank Mr. Manish Prateek for his guidance and motivation.

REFERENCES

[1] Russel, S. J., and P. Norvig. Artificial Intelligence: A Modern Approach. Prentice-Hall, Upper Saddle River, New Jersey, 2002.

Review on Time Synchronization for Wireless Sensor Networks

A. P. Zurani & B. N. Mahajan

G. H. Raisoni College of Engineering, Digdoh Hills, Wanadongari, Nagpur, India

Abstract - Wireless sensor networks consist of small devices distributed over geographical area. Each one of these devices has sensing, computing, and communicating components. Wireless sensor networks are used in many applications where partial or full time synchronization in the network is required. Time synchronization aims at equalizing the local times for all nodes in the network, if necessary. This paper explains the need for time synchronization in sensor networks , the requirements of synchronization methods for wireless sensor networks, synchronization methods for wireless sensor networks , common challenges for synchronization methods.

Keywords - *Sensor Networks, Time Synchronization, Malicious nodes.*

I. INTRODUCTION

Wireless Sensor Network (WSN) consists of hundreds or thousands of micro sensor nodes that are joining together to form a network. Wireless sensor network accurately monitors remote environment intelligently by combing the data from individual nodes. Applications of sensor networks are in providing health care for elderly, surveillance, emergency disaster relief, and battlefield intelligence gathering. Time synchronization is a critical building block in distributed wireless sensor networks. The special nature of wireless sensor network imposes challenging requirements on secure time synchronization design. Firstly, time synchronization must be highly energy efficient, since sensor nodes operate with batteries. Secondly, time synchronization must be accurate to the microsecond level as to fulfill time-critical WSN applications. Thirdly, time synchronization must be secure against passive, active, internal and external attackers.

The following sections are organized as follows: In Section II, we describe how computer clocks work and explain sources and requirements of time synchronization. Section III details main time synchronization methods; Traditional Time Synchronization(TTS), Reference Broadcast Synchronization (RBS), Timing-Sync Protocol for Sensor Networks (TPSN), Tiny-Sync and Mini-Sync, and lightweight tree-based Synchronization . Section IV concludes the paper by comparing different synchronization methods.

II. SYNCHRONIZATION ISSUES

A. Computer clocks

Computer clock circuits consist of an oscillator and a counter. Based on the oscillator angular frequency, the counter increases its value to represent the local clock C(t) of a network node. In ideal situations, angular frequency is constant. Thus, the clock rate of change dc / dt is equal to 1.

Due to physical variations, like temperature, vibration, and pressure, the angular frequency changes and computer clocks drifts. The local clock of node *i* can be related to real time *t* as follows :

$$C_i(t) = a_i t + b_i \qquad (1)$$

Where a_i is the clock *drift,* and b_i is the *offset* of node *i*'s clock. *Drift* denotes the rate (frequency) of the clock, and *offset* is the difference in value from real time *t*. Using equation 1, we can compare the local clocks of two nodes in a network, say node 1 and node 2:

$$C_1(t) = a_{12}. C_2(t) + b_{12} \qquad (2)$$

We call a_{12} the *relative drift* and b_{12} the *relative offset* between the clocks of node 1 and node 2. If two clocks are perfectly synchronized, then their relative drift is 1, meaning the clocks have the same rate. Their relative offset is zero, meaning they have the same value at that instant. Based on the previous equations, clock rate of network node and offset can be used to adjust its local time. Some schemes designed to adjust offsets of nodes repeatedly, or adjust offset and clock rate to a common time scale.

B. Sources of time synchronization errors

Message exchange is used in many time synchronization algorithms. If one node sends a packet with a time stamp, non deterministic delays like access and propagation times make it hard for the receiver node to synchronize precisely with the sender node. In general, the following elements contribute to the synchronization errors :

- *Send time:* This is the total time of building the message and transfer it to the network interface to be sent. This time highly depends on the operating systems in use.
- *Access time:* This is the time needed to access the channel. Every network employs a medium access control (MAC) scheme, like time division multiple access (TDMA), and total access time depends on that scheme. In TDMA for example, network node has to wait for its slot to start transmitting while in other schemes, network nodes wait for the channel to be idle.
- *Propagation time:* This is the time required to propagate the message through the air from network interface of the sender to the network interface of the receiver.
- *Receive time:* This is the time spent in receiving the message by network interface and transferring it to the application layer of the host.

C. The Need for synchronization in sensor networks

There are several reasons for time synchronization in sensor networks. First, sensor nodes need to coordinate their operations and collaborate to achieve a complex sensing task. Data fusion is an appropriate example of such coordination in which data collected at different nodes and aggregation of data gives a meaningful result. For example, in a vehicle tracking application, sensor nodes report the location and time that they sense the vehicle to a sink node which in turn combines this information to estimate the location and velocity of the vehicle. Clearly, if the sensor nodes lack a common timescale (i.e., they are not synchronized) the estimate will be inaccurate. Second, synchronization can be used by power saving schemes to increase network lifetime. For example, sensors may sleep (go into power-saving mode by turning off their sensors and/or transceivers) at appropriate times, and wake up when necessary. When using power-saving modes, the nodes should sleep and wake-up at coordinated times, such that the radio receiver of a node is not turned off when there is some data directed to it. This requires a precise timing between sensor nodes.

D. Requirements of Synchronization schemes for Sensor networks

When designing time synchronization algorithm, wireless network limitations enforce certain requirements that need to be met. Usually, the following metrics are used to evaluate any synchronization technique:

- *Accuracy:* Precision of synchronization technique highly depends on the application.

- *Robustness:* Network nodes might die or go out of scope because of the harsh environment they are deployed in. Any synchronization scheme should adapt to such changes in the network and function in all situations.

- *Scalability:* In some applications, tens of thousands of sensors might be deployed. Any synchronization technique must work well with any number of nodes in the network.

- *Longevity:* Based on the application, provided time synchronization may be instantaneous, e.g. when a certain event happens, or may last as long as the network operates.

- *Energy efficiency:* Network nodes have limited energy resources. All network protocols including synchronization ones should consider this limitation.

- *Cost:* Due to advanced technologies, network nodes are becoming so small and inexpensive. Any synchronization algorithm should not add cost or increase the size of network node.

- *Scope:* Time synchronization algorithm provides a common time for all nodes in the network, which costs more energy and time, or provides a common time to only spatially close nodes.

- *Delay:* Many applications, like detecting gas leak, require an immediate response. For those kinds of applications, the total required time to synchronize the network must be as low as possible.

III. SYNCHRONIZATION METHODS FOR WIRELESS SENSOR NETWORKS

Synchronization schemes aim at adjusting the local times of network nodes to the same reference value. The most strict and power consuming schemes require synchronization of all nodes in the network at all times (always-on), while other more relaxed schemes require synchronization of few nodes at a time.

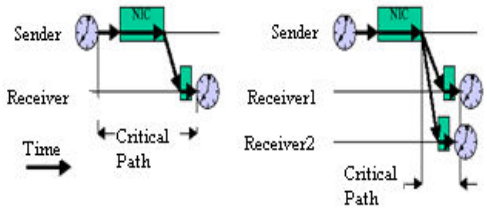


Fig. 1 : A critical path analysis for traditional time synchronization protocols (*left*) and RBS .

A. Traditional Time Synchronization (TTS)

Traditional schemes to synchronize network nodes, like Network Time Protocol (NTP), are based on sending and receiving messages. In the simplest form, sender transmits a message with its current local time to the receiver. Then the receiver adjusts its clock to the received time. This scheme works if the delay between sending and receiving messages is negligible compared to the total desired accuracy. If the delay is large, node sends a message and assuming that the receiver replies back instantaneously; node can calculate the round trip time and use that time to synchronize with the other nodes.

B. Reference Broadcast Synchronization (RBS)

RBS was introduced, instead of sender with receiver, a receiver with receiver synchronization is used. In the RBS scheme, nodes send reference beacons to their neighbors and other nodes use the arrival time of those beacons as a reference to find the time offsets between them. A reference beacon does not include a timestamp; instead, its time of arrival is used by receiving nodes as a reference point for comparing clocks.

The authors argue that by removing the sender's nondeterminism from the critical path . RBS achieves much better precision than traditional synchronization methods that use two-way message exchanges between synchronizing nodes.

A**s** the sender's nondeterminism has no effect on RBS precision, the only sources of error can be the nondeterminism in propagation time and receive time. The authors claim that a single broadcast will propagate to all receivers at essentially the same time; hence, the propagation error is negligible.

Typical communication scenario is that one node sends a beacon to its neighbors, and receivers exchange their receive time of this beacon to find their relative time offsets, and hence synchronize with each others. Precision of this scheme increases with the increase number of beacons used to synchronize.

C. Timing-Sync Protocol for Sensor Networks (TPSN)

This synchronization technique consists of two phases; level discovery phase and synchronization phase. *The* aim of the level discovery phase is to create a, hierarchical topology in the network ,where each node assigned a level; only one node is assigned level 0 and it is called the root node. In the second phase, a node of level *i* synchronizes to a node of level *i* - 1. At the end of the synchronization phase, all nodes are synchronized to the root node, and network-wide synchronization is achieved.

Level discovery phase: The first step in this level is selecting the root node. This node can be connected to an external time reference, like GPS. The root node is assigned level 0, and initiates the level discovery phase by broadcasting a *level-discovery* packet. This packet contains the identity and level of the sender node. Upon receiving this packet, the neighbors of the root node assign themselvcs level 1. Then each level 1 node broadcasts a level_discovery packet with its level and identity in the packet. Once a node is assigned a level, it discards further incoming level_discovery packets. This broadcast chain goes on through the network, and the phase is completed when all nodes are assigned a level.

Synchronization phase: Again, the root node starts the synchronization phase by sending time_sync packet. Node A of level *i* synchronizes to node B of level *i-1* through a two-way message exchange. Node A sends a packet with it is local send time T_1. Node B receives the packet at time T_2, which can be calculated as :

$$T_2 = T_1 + T_d + \Delta \qquad (3)$$

Where T_d is the propagation delay and Δ is the relative clock drift between the nodes and both assumed to be constant within the time of exchanging messages. Node B waits for a random time and responds back to node A through an acknowledgment packet at time T_3, which includes the values of T_1, T_2, T_3, and its level number. Once node A receives this packet at T_4, it can calculate Δ and T_d as follows and synchronize itself to node B.

$$\Delta = \frac{(T1-T2) - (T4-T3)}{2}, \qquad (4)$$

$$d = \frac{(T2-T1) + (T4-T3)}{2} \qquad (5)$$

This method of synchronizing all nodes in level *i* to nodes in level *i-1* continues until all nodes in the network get synchronized.

TPSN is implemented on Berkeley's Mica architecture and makes use of timestamping packets at the MAC layer in order to reduce uncertainty at the sender. TPSN achieves two times better precision than RBS. It was reported that RBS achieved a precision of

29.13 *μs,* while TPSN achieves16.9 *μs* for the same experimental setup.

D. Tiny-Sync and Mini-Sync

Tiny-Sync and Mini-Sync arc the two lightweight synchronization algorithms proposed mainly for sensor networks by Sichitiu and Vcerarittiphan.

These algorithms are extended to synchronize the whole network nodes. Both algorithms use the conventional two-way messaging Scheme to collect data points, which are used to apply tight bounds on relative drift and relative offset between two nodes. To create data point using two nodes; 1 and 2, node 1 sends a probe message with timestamp t_0. Node 2 timestamps the received message with t_b and sends back an acknowledgment to node 1, immediately or after sometime, which timestamps this acknowledgment with tr. Fig. 2 shows the previous described sequence.

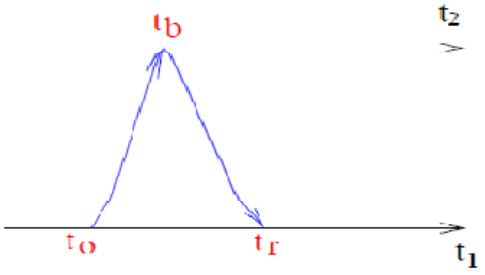


Fig. 2 : Probe message exchange

Based on equation 2 above and taking into account the order of transferred messages, the following relationships should hold :

$$t_{0(}t) < a_{12}t_b(t) + b_{12} \quad (6)$$

$$t_{r(}t) > a_{12}t_b(t) + b_{12} \quad (7)$$

Algorithms start collecting data points, using linear programming to estimate a_{12} and b_{12} , and synchronizing nodes with each others. While regular data and acknowledgment packets in the network can be used to collect data points, which reduce the overhead, estimating a_{12} and b_{12} require large storage and heavy computation.

E. Lightweight Tree-Based Synchronizafion

Lightweight Tree-Based Synchronization (LTS), proposed by Greunin and Rabaey **,** is distinguished from other work in the sense that the aim is not to maximize accuracy, hut to minimize the complexity of the synchronization. Thus, the needed sylchronization accuracy is assumed to be given as a constraint, and the target is to devise a synchronization algorithm with minimal complexity to achieve a given precision. This approach is supported by the claim of the authors that the maximum time accuracy needed in sensor networks is relatively low (within fractions of a second), *so* it is sufficient to use a relaxed, or lightweight, synchronization scheme in sensor networks.

Two LTS algorithms are proposed for multihop synchronization of the network based on pairwise synchronization scheme of, as also explained earlier. Both algorithms require nodes to synchronize to some reference point(s) such as a sink node in the sensor network. The first algorithm is a centralized algorithm, and needs a spanning tree to be constructed first. Then pairwise synchronization is done along the $i \sim 1$ edges of the spanning tree. In the centralized algorithm, the reference node is the root of the spanning tree and has the responsibility of initiating a "resynchronization" as needed. Using the assumption that the clock drifts arc hounded, and given the required precision, the reference node calculates the time period that a single synchronization step will he valid. Since the depth of the spanning tree affects the time to synchronize the whole network as well as the precision error at the leaf nodes, the depth of the tree is communicated hack to the root node ***so*** that it can use this information in its resynchronization time decision.

The second multihop LTS algorithm performs network-wide synchronization in a distributed fashion. Each node decides the time for its own synchronization, and a spanning tree structure is not used in this algorithm. When node i decides that it needs to synchronize (using the desired accuracy, its distance from the reference node, and the clock drift), it sends a synchronization request to the closest reference node (by any routing mechanism available). Then all nodes along the path from that reference node to *i* must he synchronized before node *i* can be synchronized. The advantage of this scheme is that some nodes may have less frequent events to deliver, and therefore may not need frequent synchronization.

Since nodes have the opportunity to decide on their own synchronization, this saves unnecessary synchronization effort for such nodes. On the other hand, letting each node decide on resynchronization may boost the number of pairwise synchronizations, since for each synchronization request all nodes along the path from the reference node to the resynchronization initiator need to he synchronized. As the number of synchronization requests increase, the overall effect of synchronizations along these paths may be a significant waste of resources. Hence, the idea of aggregating synchronization requests is proposed; when any node wishes to request synchronization, it queries adjacent nodes to discover the existence of any pending request. If any exists, the synchronization request of this node could be aggregated to a pending request,

decreasing the inefficiency that would be. caused by two separate synchronizations along the same path.

IV. CONCLUSION

We have presented basic synchronization methods proposed for wireless sensor networks, reviewing the motivations and requirements for such work. Two synchronization algorithms, RBS and TPSN, both report very high precisions, on the orders of few microseconds, although they use completely different approaches. The receiver-receiver synchronization of RBS completely eliminates the uncertainty at the sender, and thus is believed by many researchers to perform better than classical sender-receiver synchronization. However, it should he noted that receiver-receiver synchronization requires four messages sent and three messages received for synchronizing two nodes, while sender-receiver synchronization requires only two sent and two received messages. As radio communication is known to be the most energy. consuming component of sensor node operations, this is almost a two times increase in energy-complexity. This increase in the complexity of receiver-receiver synchronization can he reduced to some degree by synchronizing many receivers by a single synchronization pulse broadcast by the sender. Although TPSN does not suffer from energy complexity in this respect, it needs a hierarchical structure of nodes to be formed, which might increase the synchronization cost. Mini-Sync also relies on a hierarchical structure among sensor nodes, although it is a low-complexity option for synchronizing sensor networks. LTS algorithms offer very low-cost synchronization, hut with very limited accuracy and thus limited applicability.

REFERENCES

[1] M. L. Sichitiu and C. Veeroriniphon, "Simple, Accurate Time Synchronizotion for Wireless Sensor Networks," WCNC 2003.

[2] J. V. Greunen and J. Raboey, "lighttweight Time synchronization for Sensor Networks," Proc. 2ndACM lnt? C o d Wireless Sensor Networks and Apps.,San Diego, CA, Sept. 2003.

[3] K. Romer, 'Time Synchronization in Ad Hoc Networks," ACM MobiHoc '01,Lon Bwch, CA, Oct. 2001

[4] T. Haenselmann. “Sensornetworks,” Wireless Sensor Network textbook, 2006.

[5] S. Ganeriwal, R. Kumar, and M. Srivastava, “Timing Sync Protocol for Sensor Networks," ACM SenSys, Los Angeles, November 2003.

[6] J. Erson and D. Estrin, Time Synchronization for Wireless Sensor Parallel and Distrib. Comp. lssues in Wireless Netwoks Mobile Comp., San Francisco, CA. Apr. 2001.

[7] F. Sivrikaya and B. Yener, “Time Synchronization in Sensor Networks: A Survey,” Network, IEEE Volume 18, Issue 4, July- Aug. 2004, pps.45 – 50.

◈◈◈

Hybrid Channel Allocation in Wireless Cellular Networks

Shruti Pancholi & Pankaj Shukla

Electronics Engg. Dept. University College of Engg., U.C.E. Rajasthan Technical University, Kota, Rajasthan, India

Abstract - The enormous growth of mobile telephone traffic, along with the limited number of channels available, requires efficient reuse of channels. Channel allocation schemes can be divided into three categories: Fixed Channel Allocation (FCA), Dynamic Channel Allocation (DCA) and Hybrid Channel Allocation (HCA). In HCA, channels are divided into two disjoint sets: one set of channels is assigned to each cell on FCA basis, while the others are kept in a central pool for dynamic assignment. This paper presents a hybrid channel allocation notification to the central pool on each channel request that cannot be satisfied locally at the base station. This notification will request more than one channel to be assigned to the requesting cell. When FCA would not support call then borrowed channels will be in use. The simulation study of the protocol indicates that the blocking probability of HCA will be low for less traffic and high for high traffic.

Keywords - *Channel allocation, wireless networks, exponential distribution.*

I. INTRODUCTION

The recent growth of mobile telephone traffic, along with the limited number of radio frequency channels available in cellular networks, requires efficient reuse of channels. An efficient channel allocation strategy is needed and it should exploit the principle of frequency reuse to increase the availability of channels to support the maximum possible number of calls at any given time. A given frequency channel cannot be used at the same time by two cells in the system if they are within a distance called minimum channel reuse distance, because it will cause radio interference (also known as co-channel interference).Several channel allocation schemes have proposed [1] and they can be divided into three categories: Fixed Channel Allocation (FCA), Dynamic Channel Allocation (DCA), and Hybrid Channel Allocation (HCA). In FCA schemes, a fixed number of channels are assigned to each cell according to predetermined traffic demand and co-channel interference constraints. FCA schemes are very simple; however, they are inflexible, as they do not adapt to changing traffic conditions and user distribution In order to overcome these deficiencies of FCA schemes, DCA schemes have been introduced. In DCA schemes, channels are placed in a pool (usually centralized at Mobile Switching Centre (MSC) or distributed among various base stations) and are assigned to new calls as needed. Any cell can use a channel as long as the interference constraints are satisfied. After the call is over, the channel is returned back to the central pool. At the cost of higher complexity and control message overhead, DCA provides flexibility and traffic adaptability. However, DCA schemes are less efficient than FCA under high load conditions. To improve performance, some DCA schemes use channel reassignment, where on-going calls may be switched, when possible, to reduce the distance between co-channel cells. Another type of DCA strategy involves channel borrowing mechanism from neighboring cells. In such a scheme, channels are assigned to each cell as is normally done in the case of FCA. However, when a call request finds all such channel busy, a channel may be borrowed from a neighboring cell if the borrowing will not violate the co-channel interference constraints [1].

HCA techniques are designed by combining FCA and DCA schemes in an effort to take advantages of both schemes. In HCA, channels are divided into two disjoint sets: one set of channels is assigned to each cell on FCA basis (fixed set), while the others are kept in a central pool for dynamic assignment (dynamic set). The fixed set contains a number of channels that are assigned to cells as in the FCA schemes and such channels are preferred for use in their respective cells.

When a mobile host needs a channel for its call, and all the channels in the fixed set are busy, then a request from the dynamic set is made. The ratio of the number of fixed and dynamic channels plays an important role. It has been found that if the ratio of FCA and DCA forms HCA. The HCA techniques proposed in the literature are complex to implement and they suffer from the large control overhead incurred from system state collection and dissemination.

This paper presents a new HCA scheme that takes into account the level of traffic intensity and blocking probability in a cell.

II. BASIC CHANNEL ALLOCATION SCHEMES

A HCA method, which is composed of two parts. The first part is the allocation of nominal channels for each cell. This is carried out in the planning stage of wireless communication network. The second part is the allocation of channels to ongoing call requests while the wireless network is in use. This is carried out dynamically, when a call originates in a cell without free nominal channels.

Assuming a cellular structured mobile communication system layout, a point of interest is to decide on what Channel Assignment Scheme (CAS) to use. One such scheme is the Fixed Channel Assignment (FCA), where channels can only be used in designated cells [4], [6], [8], [11],[13],[14],[15] .In this case there is a definite relationship between cells and channels that can be used there at any time.

One obvious disadvantage of using this scheme can best be explained using an example. Imagine two neighboring cells with their assigned channels. If at any time, one of the cells happens to have all its channels occupied, and another request for service is made in this same cell, this new request will be denied even though there may be free channels in the neighboring cell at this very instant. The overall result is one of poor channel utilization.

Another channel assignment scheme is the Dynamic Channel Assignment scheme (DCA). In the DCA approach, there is no definite relationship between the cells of the system and the channels that are used in them. Channels are temporarily assigned for use in cells for the duration of the call. After the call is over, the channels are returned and kept in a central pool [2], [4], [6], [7], [14]. To avoid co-channel interference that would result if two neighboring cells used the same channel simultaneously, any channel that is in use in one cell can only be reassigned simultaneously to another cell in the system if the relative distance between the two cells is d, where d is defined as

$$d = D/R \qquad (1)$$

Where R is the radius of the cell and D is the physical distance between the two cell centres (the resulting average spacing between cells using the same channel depends on the criterion of borrowing, but it is usually larger than d). This physical separation that must exist between any two cells using the same channel gives rise to the concept of an Interference cell group. The interference cell group, for a given cell, is comprised of all those cells with which it can interfere if they transmit on the same channel simultaneously.

When a cell wishes to borrow a channel for temporary use, there is usually more than one channel in the central pool and therefore one has to decide which one, out of all the eligible channels, to borrow for use. Many different schemes for a cell borrowing a free channel have been investigated and published [2], [6], [7], [9], [11]. The FCA and DCA are two 'definite' policies, definite in that over the entire service area, and for all time, channels are either assigned with FCA or DCA disciplines. There are two other channel assignment schemes which are a combination of FCA and DCA. The third CAS will be called Constrained Dynamic Channel Assignment (CDCA) scheme [2]. The fourth and last CAS of interest in this paper is the Hybrid Channel Assignment (HCA) scheme [4, 15]. Explanations of these channel assignment algorithms will be given later.

FCA and DCA schemes have been studied quite extensively [6-8], [10], [11], and the results from system simulations have shown that, for low blocking probabilities, the Dynamic system performs much better than the Fixed system. But for very high blocking probabilities, which are synonymous with very large offered traffic, the FCA scheme performs better.

The initiation of requests for service from cell to cell is a random process and, therefore, when DCA is being used, the different channels are assigned to serve calls at random too. Because of this randomness, it is found that cells that have borrowed the same channel for use are, on the average, spaced apart at a greater distance than the minimum required distance d. Consequently DCA schemes are not always successful in reusing the channels the maximum possible number of times. But for FCA systems, the channel assignment to cells is done observing the minimum spacing d and, therefore, it has a higher channel reuse. This is why, in order to improve the performance of DCA systems at large traffic offerings, it has been suggested to use Channel Reassignment techniques [4]**.** The basic goal of Channel Reassignment is to switch calls already in progress, whenever possible, from channels that these calls are using, to other channels, with the objective of keeping the distance between cells using the same channel simultaneously to a minimum. It has been found that, in the case of DCA, the system is not overly sensitive to time and spatial changes in offered traffic, giving rise to almost stable grades of service in each cell [5]. But for the FCA scheme, the service deviation, a measure of the grade of service fluctuations from one cell to another, is very much worsened by these said traffic changes. Another point in favor of DCA over FCA, as deduced from simulation results, is the seeming

dependence of the grade of service within an Interference cell group on the average loading within that group and not on it spatial distribution [2], [5].A channel assignment scheme that is superior to FCA and DCA, and which will be called constrained DCA in this paper, was proposed by [9] and by [2] by comparing it to two other channel assignment schemes, using some simulation results. Concluding from results that the CDCA scheme behaved like a full access system, with the number of channels equal to the total channels available for use in the heaviest loaded interference cell group.

In this scheme, each cell has two sets of channels for its use, shown in Fig. 1 as Al, B1, C1 and (A, B), (A, C), (B, C). The former type of sets contains the nominal channels. These channels have been assigned to the cells observing the minimum interference spacing and in all cases are to be preferred for use in their respective cells (nominal cells). If all nominal channels for a particular cell are busy when a new call originates or arrives in a particular cell, then borrowing may take place from the borrow able set, shown in brackets in that cell, provided no interference will result as a consequence of this borrowing. It is of interest, to note that the set in brackets may contain many channels, and therefore the decision on which channel of the set will be borrowed is important. A general conclusion reached by most authors on this subject was that adopting a simple test for borrowing (for example, borrowing the first available channel that satisfies the d constraint) yields performance results quite comparable to systems which do a lot of exhaustive searching for channels that are the ultimate best for borrowing, thus giving rise to a lot of processing per call. Because of this reason, in the research which led to the results presented in this paper, the criterion for borrowing was simply to use the first available channel in the search.

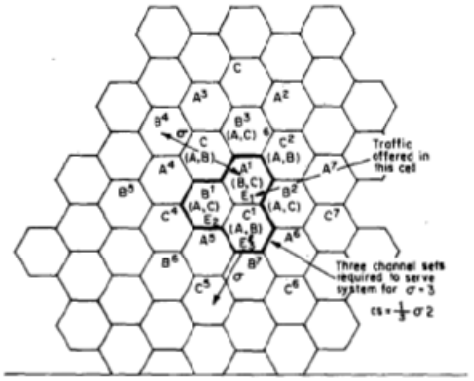


Fig. 1 : Hexagon cellular layout. Key: A^i, B^i, C^i sets of channels that are assigned to cells for use there as first choice. (A, B), (A, C),(B, C) sets of channels that can be borrowed for use in the cells where indicated, provided such borrowing meets interference constraints imposed on the system [15].

In the Hybrid Channel Assignment Scheme, employ a mixture of two schemes (thus the name Hybrid). These are the FCA and DCA schemes. Assume a total of *T* duplex channels for service and that they are divided into two sets A and B**,** not necessarily equal. Then channel set A contains channels that are used, in the system using the FCA scheme. Channel set B contains those channels that can be used in any cell in the system, using the DCA scheme [4], [15].Up to now, the question of exactly in what ratio to divide T channels into the two sets A and B has remained unanswered.

III. DESCRIPTION OF SIMULATED SYSTEM

Basic assumptions

1. For the investigation of the optimum division between fixed and dynamic channels, a system with a very large cellular layout should be used, but the statistics should be collected from the central cells only. The reason for considering a large cellular layout was to overcome the edge effects. Using a small system for this kind of study is bad because the cells around the edges do not have enough neighboring cells to cause calls to be blocked, whereas the centrally located cells have a lot of neighboring cells and therefore every time the central cells wish to borrow, chances are the neighboring cells will be using the desired channels. This gives rise to the central cells having higher blocking probabilities than those at three edges.

2. The calls in each cell are assumed to have a Poisson distribution with known arrival rate, λ calls/hour.

3. The service time per call, in any cell, is assumed to be exponentially distributed, with a mean of 180 seconds. Thus the loading will be:

$$\frac{\lambda}{3600} * 180 \quad \text{[Erlangs]} \qquad (2)$$

4. ~~The~~ first available channel in the search that satisfies the spacing constraint is borrowed for use.
3600

5. It is assumed that the mobiles are identifiable entities and could operate on any channel, as dictated by the base stations.

6. The base stations could transmit on any borrowed frequency at all times, as assigned to them by the system controller.

Now consider a system having uniform spatial offered traffic and using a HCA scheme. The steps involved in the simulation are as follows:

a) Assumed that the long term average offered traffic in Erlangs was known. Then using the tables for the Erlang B traffic formula [12] now determine the number of channels required in each cell to give the desired grade of service assuming that a FCA scheme was in use. The desired number of channels for cells 1, 2, 3 ... was represented by (NC) Fig. 2.

b) Then consider a mobile communications system with uniformly offered traffic that requires (NC) channels per cell, on the average. the ratio of Fixed to Dynamic channels that carry the most traffic at the desired grade of service. Let this, ratio be represented F: D, where D is the average number of static channels per cell, is the average number of dynamic channels per cell and F+ D = (TC).

c) Now using the results obtained in Step b above, channels are assigned to the cells of the mobile communications system. Consider, for example, a cell that has offered traffic of E, Erlangs. Then normally (TC) channels would be needed to give a desired grade of service. But from the simulation results only F channels are assigned to cell 1 and 1, channels are given to the entire system for use as Dynamic channels.

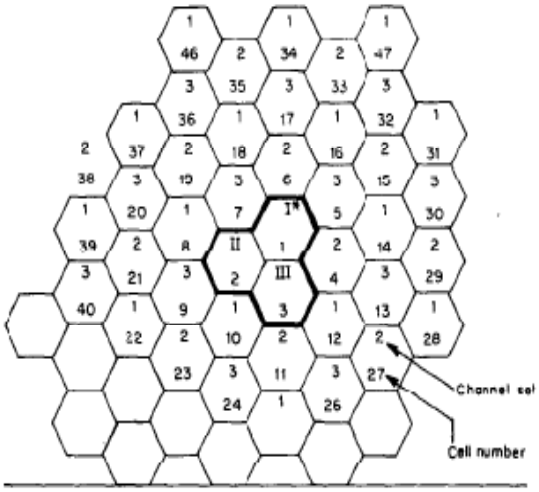


Fig. 2 : Cellular layout for system that was simulated.[15]

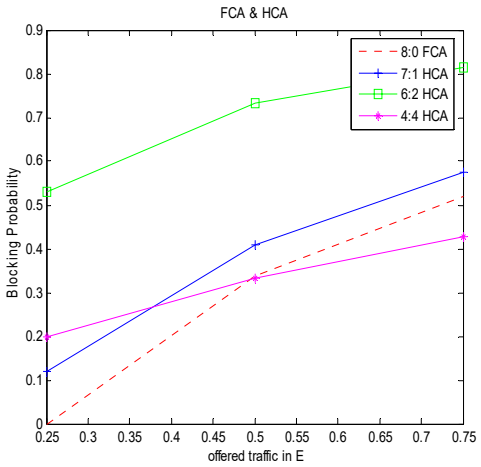


Fig. 3 : Simulation results showing the comparison between FCA and HCA

Table I : DIFFERENT SYSTEM COMBINATION INVESTIGATED

Average channels loading per cell in uniformly loaded system	Channel partition for Hybrid simulation Fixed	Dynamic	Traffic in Erlangs
8	8	0	5,6,7,8
	7	1	
	6	2	
	4	4	

IV. CONCLUSION

The obtained results indicate that the Fixed to dynamic Channel Assignment scheme ratio per cell is equal performs better than other ratios. The simulation study of the protocol indicates that the blocking probability low for less traffic in FCA case, while for high traffic fixed and dynamic channels should equal so that blocking probability remain less. Beyond this, for different combination of ratios HCA has low blocking for less traffic and high for high traffic.

To further improve the performance, there is no division of fixed and dynamic channels groups. Dynamic channel pool includes all the channels in the system.

REFERENCE

[1]. Katzela and M. Naghshineh, "Channel Assignment Schemes for Cellular Mobile Telecommunication Systems: A Comprehensive Survey", IEEE Personal Communications, vol. 3, No. 3, June 1996.

[2]. Lewis G. Anderson, "A Simulation Study of Some Dynamic Channel Assignment Algorithms in a High Capacity Mobile Telecommunications System," IEEE Trans. Veh. Technol., Vol. VT-22, pp.210-217,Nov.1973.

[3]. Kin'ichiro Araki,"Fundamental Problems of Nation-wide Mobile Radio Telephone Systems, Rev.Elec.Commun.Lab" Vol.16, pp.357-373, May-june, 1968.

[4]. Donald C. Cox .and Douglas 0. Reudink, "Increasing Channel Occupancy in Large-Scale Mobile Radio Systems: Dynamic Channel Reassignment," IEEE Trans. Veh. Technol., Vol. VT-22,pp.218-222,Nov. 1973.

[5]. "Effects of Some Non uniform Spatial Demand Profiles on Mobile Radio System Performance," IEEE Trans. Veh. Technol.,Vol. VT-21, pp. 62-67, May 1972.

[6]. "A Comparison of Some Channel Assignment Strategies in Large-Scale Mobile Communication Systems," IEEE Trans.Commun., Vol. COM-20, pp. 190-195, Apr. 1972.

[7]. "Dynamic Channel Assignment in High Capacity Mobile Communications System," Bell Syst. Tech. J., Vol. 50, pp. 1833-1857,July-Aug.1971.

[8]. "Dynamic Channel Assignment in Two-Dimensional Large-Scale Mobile Radio Systems," Bell Syst. Tech. J., Vol. 51, pp.1611-1629, Sept. 1972.

[9]. Joel S. Engel and Martin M. Peritsky, "Statistically-Optimum Dynamic Server Assignment in Systems with Interfering Servers,"IEEE Trans. Veh. Technol., Vol. VT-22, pp. 203-209, Nov. 1973.

[10]. Richard H. Frenkeil, "A High-Capacity Mobile Radiotelephone System Model Using a Coordinated Small-Zone Approach,"IEEE Trans. Veh. Technol., Vol. VT-19, pp. 172-177, May 1970.

[11]. Leonard Schiff, "Traffic Capacity of Three Types of Common-User Mobile Radio Communication Systems," IEEE Trans.Commun. Technol., Vol. COM-18, pp. 12-21, Feb. 1970.

[12]. James, Martin, Systems Analysis for Data Transmission. Prentice-Hall, Englewood Cliffs, New Jersey, 1972.

[13]. Future Developments in Telecommunications, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[14]. W. C. Jakes, Jr. (ed.), Microwave Mobile Communications, John Wiley & Sons, New York, N.Y., 1974.

[15]. T. J. Kahwa, "A Hybrid Channel Assignment Scheme in Cellular-Structured Mobile Communications Networks", M.A.Sc. Thesis, E.E. Dept., University of Ottawa, June 1977.

◈◈◈

Exploiting Data Mining Techniques For Improving the Efficiency of Time Series Data

Tarun Dhar Diwan[1], Pradeep Chouksey[2], R. S. Thakur[3] & Bharat Lodhi[4]

[1] DR. C.V.RAMAN University, Kota, Bilaspur (C.G.), India, [2] Technocrats Institute of Technology Bhopal M.P. India
[3] MANIT Bhopal M.P. India & [4]BIRT, Bhopal M.P. India

Abstract -The research work in data mining has achieved a high attraction due to the importance of its applications This paper addresses some theoretical and practical aspects on Exploiting Data Mining Techniques for Improving the Efficiency of Time Series Data using SPSS-CLEMENTINE. This paper can be helpful for an organization or individual when choosing proper software to meet their mining needs. In this paper, we propose utilizes the famous data mining software SPSS Clementine to mine the factors that affect information from various vantage points and analyse that information. However the purpose of this paper is to review the selected software for data mining for improving efficiency of time series data. Data mining techniques is the exploration and analysis of data in order to discover useful information from huge databases. So it is used to analyse a large audit data efficiently for Improving the Efficiency of Time Series Data. SPSS- Clementine is object-oriented, extended module interface, which allows users to add their own algorithms and utilities to Clementine's visual programming environment. The overall objective of this research is to develop high performance data mining algorithms and tools that will provide support required to analyse the massive data sets generated by various processes that is used for predicting time series data using SPSS- Clementine. The aim of this paper is to determine the feasibility and effectiveness of data mining techniques in time series data and produce solutions for this purpose.

Keywords *- Time series data, Data mining, Forecasting, classification, SPSS-Clementine.*

I. INTRODUCTION

Classification algorithm has discrete allowing predicting the relationship between input data sets. Commercial data mining software's are considerably expensive to purchase and the cost of training involved is high. For this, the best software that fits business needs is very important, crucial and difficult to decide the forecasting of any type of data set.

As well as modern data analysis has to cope with tremendous amounts of data. The modern economy has become more and more information-based. The widespread uses of information technology, a large number of data are collected which results in massive amounts of data. Such **time-ordered** data typically can be aggregated with an appropriate time interval, yielding a large volume of equally spaced **time series** data. Such data can be explored and analysed using many useful tools and methodologies developed in modern time series analysis.

Data mining is the exploration and analysis of data in order to discover meaningful patterns. Data mining techniques have been used to uncover hidden patterns and predict future trends. The competitive advantages achieved by data mining include increased revenue, reduced cost, and much improved marketplace responsiveness and awareness. The term data mining refers to information elicitation. It is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, data-base management, data visualization, mathematics algorithms and statistics. This technology provides different methodologies for decision-making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning and innovation.

The approach presented in this paper is a general one and can be applied to any time series data sequence. An improvement of technological process control level can be achieved by time series analysis in order to prediction of their future behaviour using SPSS-Clementine. SPSS Clementine is very suitable as a mining engine with its interface and manipulating modules that allow data exploration, manipulation and exploration of any interesting knowledge patterns. The paper deals with the utilization of data mining using SPSS-Clementine to fix best the prediction of time series. We can find an application of this prediction by the control in production of energy, heat, and etc.

SPSS Clementine software for data mining to understand Time series patterns for share marketing better curricula offerings used an classification mining tool to assist instructors in changing their pedagogical strategies and interventions by analyzing huge volumes of time series data.Classification finds patterns or a set

of models in "training" data that describe and distinguish data cases or concepts. Classification constructs a model to predict the class of objects whose class type is known. Time series data accounts for a large fraction of the data stored in financial, medical and scientific databases. Recently there has been an explosion of interest in data mining time series, with researchers attempting to index, cluster, classify and mine association rules from increasing massive sources of data. For prediction and description of time series data we are using different data mining techniques. Here prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. The second goal which leads to descriptive model, describes patterns in existing data which may be used to guide decisions as opposed to making explicit predictions.

The research work done is about the share market forecasting of SBI time series dataset. The situation for trading changes every second. A time series database consists of sequence of values or events changing with time. Time series databases are used for studying the daily fluctuation of share market.

II METHODOLOGY OF DATA MINING

A. Definition

Data mining may be defined as "the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules". Hence, it may be considered mining knowledge from large amounts of data since it involves knowledge extraction, as well as data/pattern analysis

B. Tasks

Some of the tasks suitable for the application of data mining are classification, estimation, prediction, affinity grouping, clustering, and description. Some of them are best approached in a top-down manner or hypothesis testing while others are best approached in a bottom-up manner called knowledge discovery either directed or undirected.

As for Classification, it is the most common data mining task and it consists of examining the features of a newly presented object in order to assign it to one of a predefined set of classes. While classification deals with discrete outcomes, estimation deals with continuously-valued outcomes. In real life cases, estimation is often used to perform a classification task. Prediction deals with the classification of records according to some predicted future behavior or estimated future value.

Both Affinity grouping and market basket analysis have as an objective to determine the things that can go together. Clustering aims at segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters that are not predefined. Description is concerned with describing and explaining what is going in a complicated database so as to provide a better understanding of the available data.

C. The Various Cycle of Data Mining

The four stages of the virtuous cycle of data mining are:

- Identifying the problem: where the goal is to identify areas where patterns in data have the potential of providing value.
- Using data mining techniques to transform the data into actionable information: for this purpose, the produced results need to be understood in order to make the virtuous cycle successful. Numerous pitfalls can interfere with the ability to use the results of data mining. Some of the pitfalls are bad data formats, confusing data fields, and lack of functionality. In addition, identifying the right source of data is crucial to the results of the analysis, as well as bringing the right data together on the computing system used for analysis.
- Acting on the information: where the results from data mining are acted upon then fed into the measurement stage.
- Measuring the results: this measurement provides the feedback for continuously improving results. These measurements make the virtuous cycle of data mining *virtuous.* Even though the value of measurement and continuous improvement is widely acknowledged, it is usually given less attention than it deserves.

D. DATA MINING TECHNIQUES

Data mining can be described as "making better use of data". Every human being is increasingly faced with unmanageable amounts of data; hence, data mining or knowledge discovery apparently affects all of us.

There are two different types of tools used in data mining which are classification and prediction. Classification and prediction is the process of identifying a set of common features and models that describe and distinguish data classes or concepts. The models are used to predict the class of objects whose class label is unknown. A large number of classification models have been developed for predicting future trends of stock market indices and foreign exchange rates.

Classification - is the task of generalizing known structure to apply to new data. Classification tools tend to segment data into different segments. The process of classification starts with a classification algorithm,

which is applied to a set of so called training data. The training data is fed through the classification algorithm. When the classification rules have been defined a set of non related test data can be run through the classification rules. With the result from the test data it can be estimated whether the rules work and are able to classify segments. If they show that the classification does not work to with in a desired confidence interval a new classification algorithm can be implemented to improve the results of the classification rules.The purpose of data classification is organizing and allocating data to detached classes. In this process, a primary model is established according to the distributed data. Then this model is used to classify new data. Thus, applying the obtained model, it can be determined that to which class the new data belongs.

Classification is used for discrete values and foretelling. In the process of classification, the existing data objects are classified into detached classes with partitioned characteristics (separate vessels) and are presented as a model. Then considering features of each class, the new data object is allocated to them; its label and kind becomes determinable.

In classification, the established model is obtained based on some training data (data objects that their class's label is determined and identified). The obtained model can be presented in different forms like: classification rules (If- Then), decision trees, and neural networks. Marketing, disease diagnosis, analysis of treatment effects, find breakdown in industry, credit designation and many cases related to prediction are among applications of classification.

2.1. Types of classification methods

Classification is possible through the following methods:

- Bayesian classification
- Decision trees
- Nearest neighbor
- Regression
- Genetic algorithms
- Neural networks
- Support vector machine (SVM)

Prediction- Data mining techniques provides with a level of confidence about the predicted solutions in terms of the consistency of prediction and in terms of the frequency of correct predictions. The most extensively used tools in prediction are linear and multiple regression. Linear regression is the simplest form of regression analysis where there is only one predictor variable. Where as the multiple regressions is a more complex regression analysis where there are two or more predictor variables. Also non linear regression is used in cases where there are no linear relationships with data.

Time Series: A time series is a sequence of values that a randomly varying attribute accumulates over time. A time series does not use any mechanism to adapt its values, and this makes it very different from other series. Common time series examples are stock markets, weekly weather reports, annual precipitation or weekly pizza sales. Real world time series data tend to be continuous, and are usually a sequence of observations or values separated by equal time intervals.

Time series are often presumed to consist of components that enable us to predict future patterns. These components are:

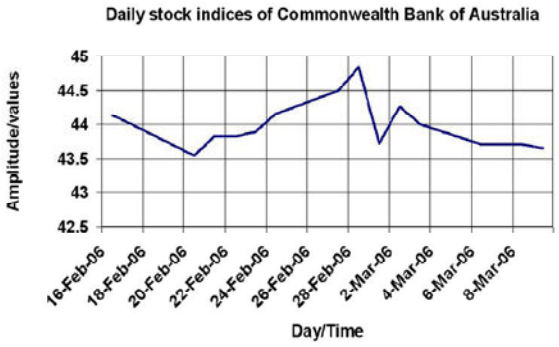
- Trend,
- Cycle,
- Seasonal variations, and
- Irregular fluctuations.

Time series prediction - Time series prediction/ forecasting is the process of studying known past events and extrapolating the results to predict future events, or in other words, the process of predicting future data points before they actually exist to confirm the measurements. Prediction of future values is complex and often difficult owing to the inherently volatile and non-linear nature of time series. Usually, prediction methods predict time series one or more steps ahead by evaluating historic data values alongside related data that may have influenced the series itself. In this thesis, we use the term prediction and forecast interchangeably to mean the forecast the future value.

Date Closing	Price
16-Feb-2006	44.13
17-Feb-2006	43.99
20-Feb-2006	43.54
21-Feb-2006	43.83
22-Feb-2006	43.82
23-Feb-2006	43.89
24-Feb-2006	44.16
27-Feb-2006	44.50
28-Feb-2006	44.85
01-Mar-2006	43.71
02-Mar-2006	44.27

03-Mar-2006	44.00
06-Mar-2006	43.70
08-Mar-2006	43.70
09-Mar-2006	43.65

Time series data: Daily stock price of Commonwealth Bank of Australia.



23-Feb-2006	43.89
24-Feb-2006	44.16
27-Feb-2006	44.50
28-Feb-2006	44.85
01-Mar-2006	43.71
02-Mar-2006	44.27
03-Mar-2006	44.00
06-Mar-2006	43.70
08-Mar-2006	43.70
09-Mar-2006	43.65

Time series data: Daily stock price of Commonwealth Bank of Australia.

2.2 Clementine software

SPSS Clementine data mining software is one of the most prominent software in data mining domain. This software is from famous SPSS software series and like previous statistical software has many facilities in data analysis domain.

The last version of this software is 12 that after its publication, next version named PASW Modeler was published. Among advantages of this software, the following cases can be mentioned:

- Consisting highly various methods for data analysis
- Very high speed in doing calculations and using database's information
- Consisting graphical environment for user's more comfort in doing analytic tasks

In new version, data cleanup and preparation are accomplished fully automatically. This software supports

All famous database software like Microsoft Office, SQL, etc.

Modules existing in this software are:

PASW Association

PASW Classification

PASW Segmentation

PASW Modeler Solution Publisher

This software can be installed on both personal computer and server; and supports 32-bit and 64-bit Windows too.

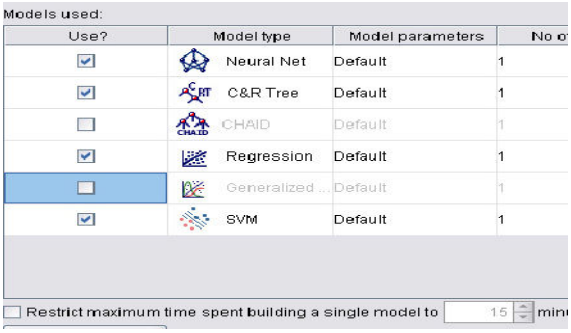
Data prepared for software

In order to make table of instructional data for classification algorithms, first we transfer column of **Max in** completely to an excel column and then transfer the same values to an opposite column in conditions that the first record is deleted.

Max in pre field is the same field that the algorithm should be finally able to predict one of them. Thus, the graph resulted from cleaned up values is presented. As it can be seen, generally, the trend of used bandwidth amount has been increasing but in some months, it has a remarkable decline.

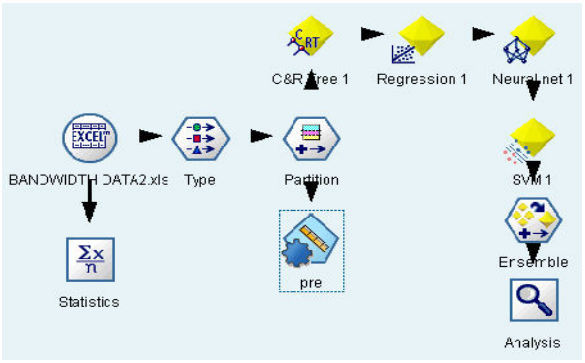
The model created in Clementine software

Now, applying Clementine software and partition node, we define eighty percent of data as instructional data and ten percent as training data and the ten remained percent as evaluation data from the final table prepared for software. Then we connect the node related to numerical prediction algorithms to related data and in its settings part, we activate the intended method.



Here, in order to implement the above algorithms, we should define columns **target** and **input**. We present column **max in** as **input** and column **max in pre** as **target**.

After implementation of algorithms and creating intended models, we create combinational model of used algorithms and compare them to each other. Figure shows the procedure for implementation of the combinational algorithm.



CLEMENTINE TOOL:

Clementine (IBM SPSS Modeler) mines useful patterns out of scattered data. It makes it easy to discover insights into your data. Its high performance increases analyst productivity. It quickly discovers patterns and trends in data more easily using a unique visual interface supported by advanced analytics.

Clementine supports various features:-

1. It allows automated data preparation- saves time when preparing the data and get to analysis phase faster.
2. Comments- document the thoughts or processes used in the creation of a model.
3. Nearest Neighbor- Quickly and easily group similar cases. Eg valuable customers or donors using prediction or segmentation techniques.
4. Statistics Integration.
5. Improved Visualization- Generate graphs from a subset of model data create rich custom graph types and style sheets seamlessly with VizDesigner.
6. It provides faster and greater return on analytical investments. Automated Modeling helps us quickly identify best-performing models and combine multiple predictions for most accurate results.
7. Proven performance and Scalable architecture- Perform data mining within existing databases and score millions of records in a matter of minutes without additional hardware requirements.
8. Improved Visualization- It enables us to visualize the breakdown of clusters, automatically generate graphs from a subset of your model data, and create custom graph types.
9. Auto cluster node- It enables users to create, sort, browse and prioritize models faster.
10. Clementine uses a visual approach to data mining that provides a tangible way to work with data. Working in Clementine is like using a visual metaphor to describe the world of data, statistics and complex algorithms.

Applications of CLEMENTINE TOOL: -

Clementine is used to mine vast repositories of data. It offers a strategic approach to finding useful relationships in large data sets.

1) **Public Sector**:- Governments around the world use data-mining to explore massive data- stores, improve citizen relationships, detect occurrences of frauds example money-laundering and tax evasion, detect crime and terrorist patterns and enhancing the expanding realm of e-govt.

2) Drug discovery and Bioinformatics: Data mining aids both pharmaceutical and genomics research by analyzing the vast data- stores resulting from increased lab automation. Clementine□s clustering and classifications models help generate leads from compound libraries while sequence detection aids the discovery of patterns.

3) Web Mining:- With powerful sequencing and prediction algorithms, Clementine contains the necessary tools to discover exactly what guests do at a Web site and deliver exactly the products or information they desire. From data preparation to modeling, the entire data mining process can be managed inside of Clementine.

4) Clementine provides templates for many data mining applications. Clementine Application

Templates CATs are available for the following activities web mining, fraud-detection, Analytical CRM, Micro array analysis, crime detection and prevention, etc.

5) Customer Relationship Management: CRM can be improved thanks to smart classification of customer types and accurate predictions of churn. Clementine has successfully helped businesses attract and retain the most valuable customers in a variety of industries.

2.3 USAGE OF CLEMENTINE IN VARIOUS PHASES OF DATA MINING

Data Mining offers a strategic approach to finding useful relationships in large data sets. In contrast to traditional statistical methods, you do not need to know what you are looking for. You can explore your data, fitting different models and investigate different relationships until you find useful information. There are various phases of data mining in which Clementine can help.

VISUALIZATION:-Clementine helps us gain an overall picture of our data. We can create plots and charts to explore relationships among the fields in our data set and generate hypotheses to explore during modeling.

MANIPULATION: - It lets you clean and prepare the data for modeling. We can sort or aggregate data, filter out fields, discard or replace missing values and derive new fields.

MODELING: - It gives us the broadest range of insight into the relationships among data fields.

Models perform a variety of tasks e.g. predict outcomes, detect sequences and group similarities.

III . RESULT AND DISCUSSION

In this section, CRISP-DM methodology has been implemented on real data set. Focus of this research is on Data Preparation and Modeling phases of this Standard for a Classification problem in order to SBI share market data set. The followings are the explanation for both of these steps:

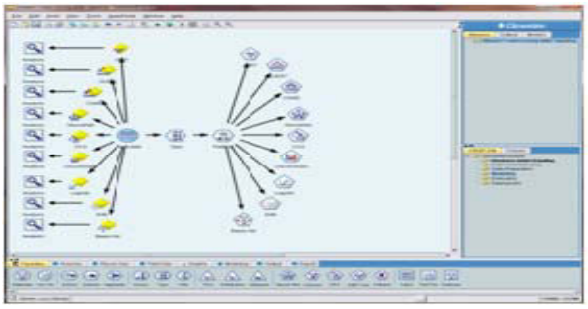
A. Data Preparation Step:

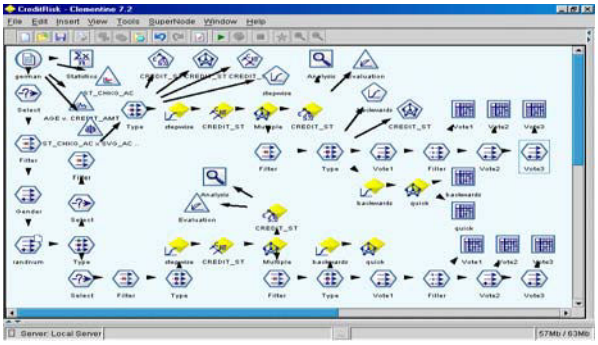
The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool) from the initial raw data. At First, the researchers have removed the noisy data from all the available records. The description of the data fields and the sample data gives us detailed information about the experimental data.

Modeling Step:

In this phase, various modeling techniques are available to be applied and their parameters are calibrated to optimal values.

Here, we have done the Classification Data Mining Algorithms with the Clementine tool on these data. Clementine data stream is shown in Fig. **Clementine Data Stream for Implementing Classification Algorithms**





Data Mining Diagram

B. Data Collection

Sources of Data

-The data set is taken for forecasting the future values consists of the financial data i.e., stock market data. The historical data set for share market gives the trends and seasonality patterns that help us to decide the accurate model for forecasting the future values and thus helps the investors to make better decision to buy or sell the share to gain profit in their Business.
We can compare the results with the ones from other models as well. Having the good results with this Hybrid model is not so important in this survey and to find a solution to forecast the problems that the output field (Class field) is being determined by multi criteria, is a proper work and overcomes the exist challenges and vagueness.

Table 1:- Sample Data Table

Record#	Date	Open	High	Low	Close	Volume
1	2011-05-02	2811.5	2819.5	2320.0	2327.6	798900.0
2	2011-04-01	2772.0	2959.9	2707.0	2805.6	349000.0
3	2011-03-01	2651.0	2888.0	2523.55	2767.9	445000.0
4	2011-02-01	2651.9	2813.4	2478.6	2632.0	662100.0
5	2011-01-03	2830.0	2852.4	2468.8	2641.0	819700.0
6	2010-12-01	2998.0	3172.0	2655.7	2811.0	682500.0
7	2010-11-01	3187.0	3515.0	2777.0	2994.1	694200.0
8	2010-10-01	3250.0	3322.0	3077.0	3151.2	280700.0
9	2010-09-01	2772.0	3268.0	2738.75	3233.2	478900.0
10	2010-08-02	2520.0	2884.0	2511.0	2764.8	482200.0

11	2010-07-01	2290.0	2519.9	2254.4	2503.8	343500.0
12	2010-06-01	2260.0	2402.5	2201.0	2302.1	403300.0
13	2010-05-03	2291.0	2348.8	2138.0	2268.3	505800.0
14	2010-04-01	2085.0	2318.8	2015.0	2297.9	452200.0
15	2010-03-02	1990.0	2120.0	1978.0	2079.0	357400.0
16	2010-02-01	2045.0	2059.9	1863.0	1975.8	550300.0
17	2010-01-04	2265.0	2315.2	1957.0	2058.0	583500.0
18	2009-12-01	2253.0	2374.7	2126.2	2269.4	585800.0
19	2009-11-03	2190.0	2394.0	2059.1	2238.1	725500.0
20	2009-10-01	2180.1	2500.0	2048.2	2191.0	855900.0
21	2009-09-01	1760.0	2235.0	1710.1	2195.7	508000.0
22	2009-08-03	1825.0	1886.9	1670.0	1743.0	428700.0
23	2009-07-01	1737.9	1840.0	1512.0	1814.0	650300.0
24	2009-06-01	1875.0	1935.0	1612.0	1742.0	651200.0
25	2009-05-04	1300.0	1891.0	1225.0	1869.1	924500.0
26	2009-04-01	1079.7	1355.0	980.0	1277.7	1100200.0
27	2009-03-02	1010.0	1132.2	894.0	1066.5	1214400.0
28	2009-02-02	1141.0	1205.9	1008.3	1027.1	814100.0
29	2009-01-01	1294.4	1376.4	1031.0	1152.2	1106800.0
30	2008-12-01	1095.0	1325.0	995.0	1288.2	1504000.0
31	2008-11-03	1155.0	1375.0	1025.0	1086.8	1426500.0
32	2008-10-01	1480.0	1569.9	991.1	1109.5	1220600.0
33	2008-09-01	1376.0	1618.0	1353.0	1465.6	993100.0
34	2008-08-03	1396.0	1638.9	1302.0	1403.6	840500.0
35	2008-07-01	1120.0	1567.5	1007.0	1414.7	674000.0
36	2008-06-02	1450.0	1496.7	1101.1	1111.4	474100.0
37	2008-05-02	1796.0	1840.0	1438.2	1443.3	368500.0
38	2008-04-01	1611.0	1819.9	1592.0	1776.3	369400.0

The data consist of historical data i.e. Stock market related fields such as open, high, low, close, volume and adjacent close and date field. The sample data is helpful in analysing the overall data set. It shows the various fields used in the dataset as well as the time interval at which the data are recorded.

C. Data Cleaning

-Real world data, like data acquired, tend to be incomplete, noisy and inconsistent. Data cleaning routines attempt to fill on missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

1) Missing Values

Many methods were applied to solve this issue depending on the importance of the missing value and its relation to the search domain.

- Fill in the missing value manually
- Use a global constant to fill in the missing value

2) Noisy Data

Noise is a random error or variance in a measured variable. Many techniques were used to smooth out the data and remove the noise.

• Clustering

Outliers were detected by clustering, where similar values are organized into groups, or clusters, values that fall outside of the set of clusters may be considered outliers.

• Combined computer and human inspection

Using clustering techniques and constructing groups of data sets, human can then sort through the patterns in the list to identify the actual garbage ones. This is much faster than having to manually search through the entire database.

3) Inconsistent Data

There may be inconsistencies in the data recorded for some transactions. Some data inconsistency may be corrected manually using external references, for example errors made at data entry may be corrected by performing a paper trace (the most used technique in our search, to guarantee the maximum data quality possible, by reducing prediction factors). Other inconsistency forms are due to data integration, where a given attribute can have different names in different databases. Redundancies may also exist.

D. Data Integration

Data Mining often requires data integration, the merging of data from multiple data sources into one coherent data store.

Careful integration of the data from multiple sources helped reducing and avoiding redundancies and inconsistencies in the resulting data set. This helped improving the accuracy and speed of the subsequent mining process.

E. Data Selection

Selecting fields of data of special interest for the search domain is the best way to obtain results relevant to the search criteria. We carefully selected from the overall data sets, and mining techniques were applied to these specific data groups in order to reduce the interesting patterns reached to the ones that represent an interest for the domain.

F. Data Transformation

In Data Transformation, the data is transformed or consolidated into forms appropriate for mining.

• **Smoothing:** which works to remove the noise form data. Such techniques include binning, clustering, and regression.

• **Aggregation:** where summary or aggregation operations are applied to the data.

• **Generalization of the data:** where low-level data are replaced by higher-level concepts through concept hierarchies.

• **Normalization:** where the attribute data are scaled so as to fall within a small specified range.

• **Attribute construction:** where new attributes are constructed and added from the given set of attributes to help the mining process.

G. Data Mining

1) Choosing the Tool

SPSS Clementine 8.1

As a data mining application, Clementine offers a strategic approach to finding useful relationships in large data sets. In contrast to more traditional statistical methods, you do not necessarily need to know what you are looking for when you start. You can explore your data, fitting different models and investigating different relationships, until you find useful information.

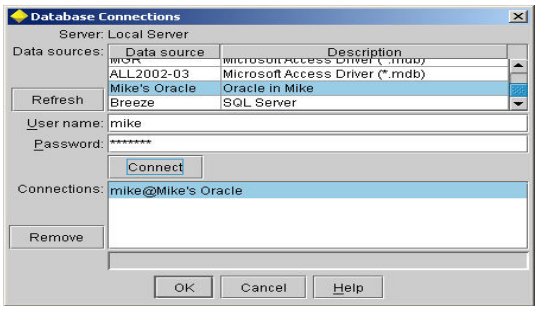
Working in Clementine is working with data. In its simplest form, working with Clementine is a three-step process. First, you read data into Clementine, then run the data through a series of manipulations, and finally send the data to a destination. This sequence of operations is known as a **data stream** because the data flows record by record from the source through each manipulation and, finally, to the destination--either a model or type of data output. Most of your work in Clementine will involve creating and modifying data streams.

At each point in the data mining process, Clementine's visual interface invites your specific business expertise. Modeling algorithms, such as prediction, classification, segmentation, and association detection, ensures powerful and accurate models. Model results can easily be deployed and read into databases, SPSS, and a wide variety of other applications. You can also use the add-on component, Clementine Solution Publisher, to deploy entire data streams that read data into a model and deploy results without a full version of Clementine. This brings important data closer to decision makers who need it. The numerous features of Clementine's data mining workbench are integrated by a visual programming interface. You can use this interface to draw diagrams of data operations relevant to your business. Each operation is represented by an icon or node, and the nodes are linked together in a stream representing the flow of data through each operation.

2) Using the Tool

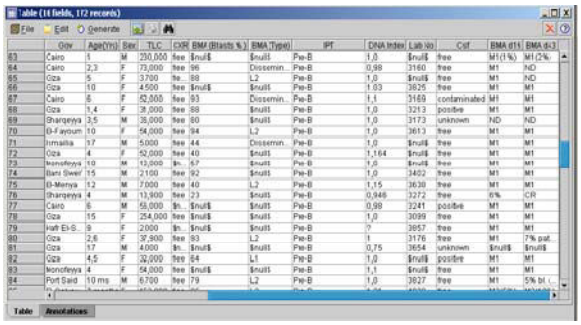
Creating the Data Source to use:

In our case an Oracle 9i Database was set and data fed into it in one table, ODBC was used to link the data source with the Clementine engine.



Viewing the Data in a Tabular Form

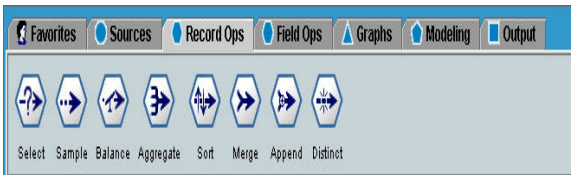
After linking the software with the data source, we view the data from the database in a tabular form by linking the data source to a "table" output



This enables us to assure that the link is successfully built and let us take a look on the form of data read by the software to detect any loss, inconsistency or noise that may have occurred in the linking process.

Manipulating the Data

By using the record Ops, operation concerning the records as a whole can be applied and used to operate on the data, using sampling, aggregation, sorting etc



Record Ops are linked to the data source directly and their output can be in any "output" means or can be directly fed as an input to other functions. Using the Field Ops on specific fields allows us to explore the data deeper, by using type selecting and filtering some fields as input or output fields, deriving new fields and binning fields.

H. Data Evaluation

After applying the data mining techniques comes the job of identifying the obtained results, in form of interesting patterns representing knowledge depending on interestingness measures. These measures are essential for the efficient discovery of patterns of value to the given user. Such measures can be used after the data mining step in order to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones. More importantly, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy pre-specified interestingness constraints.

EMAMINING THE MODEL:

We can compare the testing data set for the months Jan 2011, Feb 2011, Mar 2011, Apr 2011, and May 2011, for both Expert Modeller and Exponential Smoothing. We analysed to know which model forecast better. The one, which gives better result and is more acceptable, will used to forecast the share values or the coming five months.

Table 2:- Time series values for validation

Record#	Model	Date	STS Forecasted values for Testing				
			open	high	Low	Close	volume
97	Expert modeller	Jan 2011	2990.1	3264.4	2748.4	2653.8	733958.8
98		Feb 2011	2909.8	3359.5	2844.4	2573.6	777626.7
99		Mar 2011	2869.5	3457.3	2943.8	2520.6	814683.1
100		Apr 2011	2825.2	3558.0	3046.6	2718.0	846129.1
101		May 2011	2906.7	3661.7	3153.0	2801.5	872814.1
97	Exponential Smoothing	Jan 2011	2990.1	3134.5	2572.4	2699.0	665905.1
98		Feb 2011	2909.8	3043.3	2558.1	2661.2	769676.0
99		Mar 2011	2869.5	3012.2	2470.6	2614.3	598313.2
100		Apr 2011	2825.2	3036.0	2516.8	2690.4	330289.0
101		May 2011	2906.7	3142.0	2526.2	2741.4	523515.1

Table 3: Statistical Measures

odel	Attributes	Stationary R**2	Q	df	Sig
Expert modeller	open	0.555	21.83	15.0	0.112
	high	0.202	19.66	18.0	0.352
	Low	0.13	18.93	18.0	0.396
	Close	0.633	46.33	17.0	0.5
	volume	0.761	23.46	17.0	0.135
Exponential Smoothing	open	0.555	21.83	15.0	0.112
	high	0.527	17.37	15.0	0.297
	Low	0.536	28.86	15.0	0.017
	Close	0.562	20.72	15.0	0.146
	volume	0.677	27.56	15.0	0.024

This statistics provides an estimate of the proportion of the total variation in the series that is explained by the model. The higher value (to a maximum of 1.0), the better the fit of the model.
From the statistical measures and autocorrelation function/partial autocorrelation function we must conclude that the expert modeller will give better result as compared to exponential smoothing. It forecast the future values for the time series interval with more specificity and accuracy. Thus, for forecasting the share values for the coming five months we can use the expert modeller techniques.

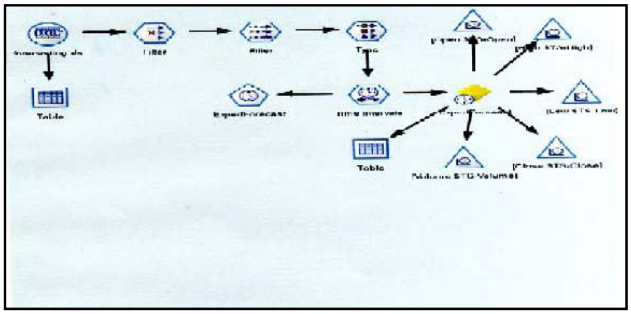
I. Knowledge Representation (outcome)

In this step visualization and knowledge representation techniques are used to present the mined knowledge to the user. All the operations applied on the records and fields, and the mining processes itself are represented in the form visualizations and graphics in this step.

STREAM MODELING FOR FORECASTING DATASET:

The stream designed for forecasting the data set is shown below:

STREAM MODEL FOR FORECASTING



The stream designing and source data is same. The source node for forecasting includes the monthly data from Jan 2003 to May 2011.

The expert modeller method is used for forecasting. The expert modeller method is set in the time series node. The expert modeller automatically select winters additive model or open & close fields; while ARIMA model is chosen for high, low, and volume fields. The golden executable nugget is connected with a table for showing the generated time series forecasted values.

Table 4:- **Forecasted Values**

Record #	Model	Date	Forecasted values				
			open	high	Low	Close	volume
102.	Expert modeller	Jun2011	2816.9	2792.5	2394.1	2299.7	1022627.1
103.		Jul 2011	2768.3	2813.8	2470.6	2415.0	869841.9
104.		Aug2011	2887.3	2835.2	2549.6	2450.5	1018116.7
105.		Sep2011	2926.8	2856.8	2631.1	2651.5	916858.6
106.		Oct2011	3133.0	2878.6	2715.2	2607.7	1015127.5

Model	Attributes	Stationary R**2	Q	df	Sig
Expert modeller	open	0.546	21.212	15.0	0.13
	high	0.019	7.599	17.0	0.974
	Low	0.124	17.513	18.0	0.488
	Close	0.519	20.421	15.0	0.159
	volume	0.798	15.372	16.0	0.498

Table 5:- **Statistical Measures**

Table 4 gives the time series forecasted values of share market for open, high, low, close, and volume having record numbers from 102 to 106. Table 5 show the various statistical error measures.

IV. CONCLUSIONS

In this paper, with the use of Time series data, we assigned the Class to the records of a database and then the Data Mining algorithms have been done on the records with Clementine software. Correct Labeling of records is so essential in Data Mining, otherwise the Data Mining is not an integrated solution and we can not trust on the results.

Based on the previous work, the following conclusions were drawn:

1. Classification, as a data mining technique, is very useful in the process of knowledge discovery in the share market field, especially in the domains where available data have many limitations like inconsistent and missing values.

   In addition, using this technique is very convenient since the Classification is simple to understand, works with mixed data types, models non-linear functions, and most of the readily available tools use it.

2. SPSS Clementine is very suitable as a mining engine with its interface and manipulating modules that allow data exploration, manipulation and exploration of any interesting knowledge patterns

3. Using better quality of data influences the whole process of knowledge discovery, takes less time in cleaning and integration, and assures better results from the mining process.

4. Using the same data sets with different mining techniques and comparing results of each technique in order to construct a full view of the resulted patterns and levels of accuracy of each technique may be very useful for this application.

So data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of time series data using SPSS- Clementine. However, data mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved.We conducted an extensive consolidation on representation methods for time series data.

The application of data mining is extremely important. It is conclusive that the average error for simulations using lots of data is smaller than that using less amount of data. That is more data for training gives better prediction. If the training error is low, predicted values are close to the real values.

ACKNOWLEDGMENT

This work is supported by research grant from MPCST, Bhopal M.P., India, Endt.No. 2427/CST/R&D/2011 dated 22/09/2011.

REFERENCES

[1] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, CA, 2005.

[2] Lon-Mu Liu, Siddhartha Bhattacharyya, Stanley L. Sclove, Rong Chen (*) etal has paper DATA MINING ON TIME SERIES: AN ILLUSTRATION USING FAST-FOOD RESTAURANT FRANCHISE DATA (1-28).

[3] G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). Time Series Analysis: Forecasting and Control. Third Edition. Prentice Hall.

[4] Chaudhuri, S. and Dayal, U. (1997). "An Overview of Data Warehousing and OLAP Technology" ACM SIGMOD Record 26(1), March 1997.

[5] Fayyad, U. M. (1997). "Editorial." Data Mining and Knowledge Discovery 1: 5-10.

[6] Friedman, J. H. (1997). "Data Mining and Statistics: What's the Connection?" Proceedings of Computer Science and Statistics: the 29th Symposium on the Interface.

[7] E. J. Keogh. A Decade of Progress in Indexing and Mining Large Time Series Databases. In VLDB, 2006.

[8] E. J. Keogh and S. Kasetty. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. Data Min. Knowl. Discov. 7(4), 2003.

[9] Weiss, S. M. and Indurkhya, N. (1998). Predictive Data Mining. San Francisco: Morgan Kaufmann Publishers. Widom, J. (1995). "Research Problems in Data Warehousing". Proceedings of 4th International Conference on Information and Knowledge Management (CIKM), November 1995

[10] http://www.jatit.org/volumes/research-papers/Vol4No12/1Vol4No12.pdf,

[11] Proceedings of the 5th National Conference; INDIACom-2011 Copy Right © INDIACom-2011 ISSN 0973-7529 ISBN 978-93-80544-00-7.

[12] IJCSNS International Journal of Computer Science and Network Security, Vol.11, No.6, June 2011 262

[13] International Journal of Information and Education Technology, Vol. 1, No. 2, June 2011.

[14] World Academy of Science, Engineering and Technology 8 2005 309 Spring/Summer 2007.

[15] Computing For Nation Development, March 10 – 11, 2011

[16] Bharati Vidyapeeth□s Institute of Computer Applications and Management, New Delhi.,

[17] Data Mining – Decision Tree Induction in SAS Enterprise Miner and SPSS Clementine – Comparative Analysis Zulma Ramirez 2901 N Juan St. Edinburg, TX 78541 (956)802-6283

◈◈◈

Application of Data Mining Technique in Stock Market : An Analysis

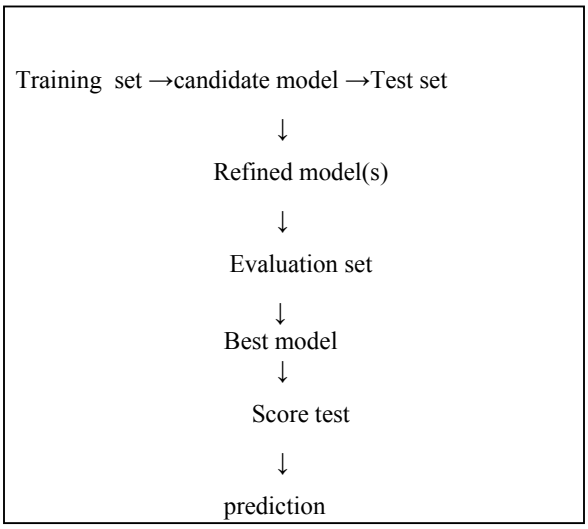
[1]Sachin Kambey, [2]R. S. Thakur & [3]Shailesh Jalori
[1]Department of Computer Applications, SATI Vidisha
[2]Department of Computer Applications, MANIT, Bhopal (MP), India
[3]Department of applied Maths & Computer science, SATI, Vidisha

Abstract -Stock market prediction with data mining technique is one of the most important issues to be investigated and it is one of the fascinating issues of stock market research over the past decade. Many attempts have been made to predict stock market data using statistical and traditional methods, but these methods are no longer adequate for analyzing this huge amount of data. Data mining is one of most important powerful information technology tool in today's competitive business world, it is able to uncover hidden patterns and predict future trends and behavior in stock market. This paper also highlights the application of association rule in stock market and their future movement direction.

Keywords - *Stock market, data mining, association rules.*

I. INTRODUCTION

Data mining, the science and technology of exploring data in order to discover unknown Patterns, is a part of the overall process of knowledge discovery in databases (KDD) [1]. In today's computer driven world, these databases contain quantities of information, exploration of this information makes data mining a matter a considerable importance and necessity. Stock market produces huge datasets that deals enormously complex and dynamic problems with data mining tool. Potential significant benefits of solving these problems motivated extensive research for years [2]. The research in data mining has gained a high attraction due to the importance of its applications and increasing generated information.A stock market or equity market is a private or public market for the trading of company stock and derivates of company stock at an agreed price; there are securities listed on a stock exchange as well as those only traded privately [4]. The expression "stock market" refers to the market that enables the trading of company stocks collective shares, other securities and derivatives. These stocks are listed and traded on stock exchanges which are entities a corporation or mutual organization specialized in the business of bringing buyers and sellers of stocks and securities together [5].Generally, data mining (some times called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relation databases [6]. Predicting the market behavior from stock database is very difficult and challenging because stock prices are dynamic. There are various steps are used for building the predictive model.They are:-



II. CHALLENGES OF STOCK MARKET

Data mining is the emerging methodology used in stock market, finding efficient ways to summarize and visualize the stock market data to give individuals or

institutions useful information about the market behavior for investment decision [7]. The enormous amount of valuable data generated by stock market has attracted researchers to explore this problem domain using different methodologies. Stock market contains various challenges which are:

1. Scientific research that relates to creation of knowledge from stock market data set.
2. Better Stock price prediction that concerns with the purchasing and sale of the items.
3. To develop feasible and efficient methods for finding the useful patterns and future trends.
4. To utilize the capital resources of the investors.
5. To boost the economy.
6. To create the interests in the favor of the stock market.
7. To protect investors and investment.
8. To maintain market stability.
9. To check out the all fraud and illegal trade practices like Harshad Mehta share scandal, Satyam scandal etc.

III. APPLICATION OF ASSOCIATION RULES IN STOCK MARKETS

As stated in Agrawal (1993) discovering association rules is an important data mining problem and there has been considerable research on using association rules in the field of data mining problem. The association's rules algorithm is used mainly to determine the relationships between items or features that occur synchronously in the database. For instance, if people who buy item x also buy item y and this information is useful for decision makers. Therefore, the main purpose of implementing the association rules algorithm is to find synchronous relationships by analyzing the random data and to use these relationships as a reference during decision making [7].One of the most important problems in modern finance is finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information about the market behaviors for investment decision [8]. The enormous amount of valuable data generated by the stock market has attracted researcher to explore this problem using different methodologies [3]. Aurangzeb Khan (2009) investigated stock market investment issues on Taiwan stock market data using a two stage data mining approach. The first stage apriori algorithm is used to mine knowledge and illustrate patterns and rules in order to propose stock category association and possible stock category investment collections. Then second stage k-means clustering algorithm is used to explore the stock cluster in order to mine stock category clusters for investment information [9].

IV. CONCLUSION AND FUTURE WORKS

With the increase of economic globalization and evolution of information technology, financial data are being generated and accumulated at an unprecedented pace. As a result, there has been a critical need for automated approaches to effective and efficient utilization of massive amount of financial data to support companies and individuals in strategic planning and investment decisions making. Data mining have been used to uncover hidden patterns and predict future trends and behaviors in financial markets. This paper we have employed the data mining and its application in stock market. Next time it is hoped that more interesting results will follow further exploration of stock data.

REFERENCES

[1] Connolly T. C. Begg and A. Strachan (1999) Database Systems: A practical Approach to Design, Implementation, and Management (3rd Ed.). Harlow: Addison-Wesley. 687, 1999.

[2] L. k. Soon and Sang Ho Lee "Explorative Data Mining on Stock Data Experimental Results and Findings", Pringer- ADMA 2007, Lnai 4632, PP. 562-569, 2007.

[3] Ehsan Hazjizadeh, Hamed Dawari Ardakani and Jamal Shahrabi", Application of Data Mining Techniques in Stock Market", journal of Economics and International Finance Vol. 2(7), PP. 109-118, July2010.

[4] Stock Exchange "IJCSNS" International Journal of Computer Science and Network Security, VOL. 7 No. 12, China 2008.", University Technology PETRONAS, 2009.

[5] Sander J., Ester M. "Techniques and Software for Development and Evaluation of Trading Strategies", International Journal of Advanced Computer Science and Applications- IJACSA, 1(5) 22-25, 2011.

[6] A Web Mining Project on Stock Market Predictor by Aaron Stone, Matthew Sullivan, Abel Canamar, 2003.

[7] Agrawal R., Imilienski T.Swami A (1993) Mining Association Rules between Set of Items in Large Databases. In proceedings of the ACM SIGMOD International Conference on Management of Data.

[8] Han and Kamber "Data Mining Concepts and Techniques ", 2nd Edition, p-234.

[9] Aurangzeb Khan, Khairullah Khan "Frequent Patterns Mining of Stock Data Using Hybrid Clustering Association Algorithm

◈◈◈

A Novel Approach For Information Security With Automatic Variable Key Using Fibonacci Q-Matrix

Shaligram Prajapat, Amber Jain & Ramjeevan Singh Thakur

International Institute of Professional Studies, D. A. University, Indore (MP), India
Computer Applications Department MANIT, Bhopal (MP), India

Abstract - **I**nformation security is essential nowadays. Large number of cipher generation and decryption algorithms exists and are being evolved due to increasing demand of users and e-commerce services. In this paper we propose a new approach for secure information transmission over communication channel with key variability concept in symmetric key algorithms using Fibonacci Q-matrix. Proposed approach will not only enhance the security of information but also saves computation time and reduces power requirements that will find it's suitability for future hand held devices and online transaction processing.

Keywords-*cipher; key; encryption; decryption; fibonacci; Q- matrix;, symmetric key algorithm, automatic variable key*.

I. INTRODUCTION

Information security plays a pivotal role nowadays. The requirement of information security is increasing because of widespread use of distributed systems, network and communication facilities for carrying information between terminal user and computer and between computer and computer [1]. Hence to provide confidentiality authentication, integrity and non-repudiation, information security has evolved.

Large number of algorithms and techniques are designed for secure transmission of data. Cryptographic algorithms play a central role in information security systems. There are two general types of key-based algorithms: Symmetric and Asymmetric algorithms. *Symmetric algorithms* (also called secret-key algorithms) are algorithms where the encryption key can be calculated from the decryption key and vice versa. In most symmetric algorithms, the encryption and decryption key are the same. Both the sender and receiver agree on a key before they can communicate securely. On the other hand, in *Asymmetric algorithms* (also called public-key algorithms) the decryption key cannot be calculated from the encryption key. So, keys play an important role in the security of any cryptographic algorithm. If weak key is used in algorithm, then any intruder may decrypt the data. One of the central factors contributing to the strength of symmetric key algorithms is the size of key used. In practice, most state-of-art cryptographic algorithms rely on increasing the key size to strengthen the security of algorithm [2]. In this paper, we instead focus for power efficient and fast algorithm based on varying the key to increase the security of algorithm.

II. RELATED WORK

Symmetric algorithms can be divided into two types: Block ciphers and Stream ciphers. *Block cipher* processes the input one block of element at a time, producing an output block for each input block. *Stream ciphers* process the input element continuously, producing output one element at a time, as it goes along. In [1, 2, 3, 6], various cryptographic algorithms [see fig. 1] and their applications have been defined and discussed.

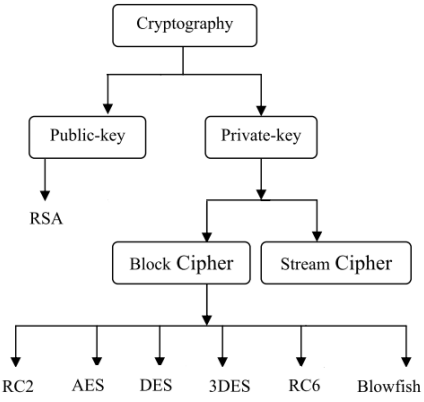


Figure 1: Overview of field of cryptography

Some symmetric block cipher algorithms are summarized in the Table 1. Literature survey reveals

that cryptographic algorithms are improving with time. Asymmetric algorithms are almost 1000 times slower than symmetric algorithms, because they require more computational processing power [4].

Table 1: Summary of some symmetric block cipher algorithms

S. No.	Algorithm	Block size	Key length
1	DES	64 bits	56 bits
2	3DES	64 bits	168, 112 or 56 bits
3	RC2	64 bits	8-128 bits (variable length key)
4	Blowfish	64 bits	32-448 bits (variable length key)
5	AES	128 bits	128, 192 or 256 bits
6	RC6	128 bits	128, 192 or 256 bits
7	RSA	-	1024-2048 bits (variable key length)

A study was performed for analyzing the performance of security algorithms by varying the key size. The effect of changing the key size on power consumption in shown in Fig. 2 and Fig. 3 [5].

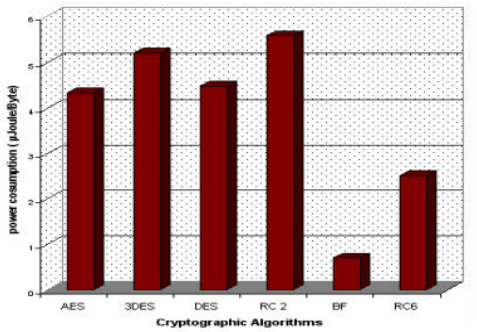


Fig. 2: Power consumption to encrypt different text document file [5]

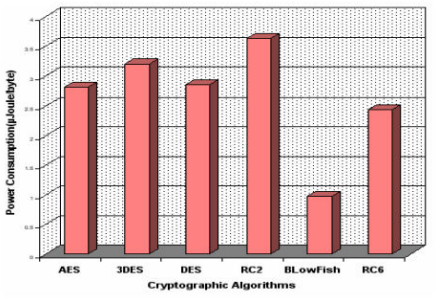


Fig. 3: Power consumption to encrypt different text document files [5]

The effect of changing the key size of AES (symmetric algorithm) on computation time in shown in Fig. 4.

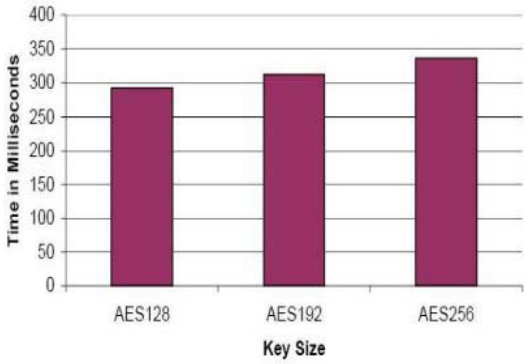


Figure 4: Time consumption for different key size for AES [4]

The effect of changing the key size of RC6 (symmetric algorithm) on computation time in shown in Fig. 5.

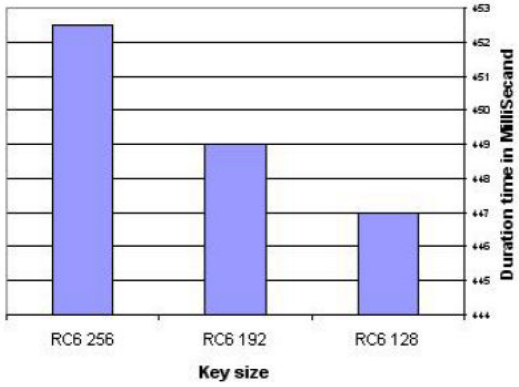


Figure 5: Time consumption for different key size for RC6 [4]

It was shown that larger key sizes lead to increase in computation time and battery power consumption.

Reversible functions are backbone of symmetric key algorithms. Many well known symmetric algorithm have been proposed using reversible *XOR* function. Stakhov proposed a coding/decoding system based on Fibonacci Q-matrix [8]. The Q-matrix is based on following concepts:

A) Fibonacci-Number

The Fibonacci numbers are obtained by following recursive function:

$$F_n = n \qquad \text{if } n = 0 \text{ or } n = 1$$

$$F_n = F_{n-1} + F_{n-2} \quad \text{if } n > 1 \qquad (1)$$

B). Fibonacci Q-Matrix

$$Q = \begin{bmatrix} F_2 & F_1 \\ F_1 & F_0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \qquad (2)$$

where Det(Q) = –1.

The nth power of this Q-Matrix can be computed as follows:

$$Q^n = \begin{bmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{bmatrix} \qquad (3)$$

where F_{n-1}, F_n and F_{n+1} are Fibonacci numbers.

Since $\text{Det}(A^n) = (\text{Det } A)^n$

Therefore, $\text{Det}(Q^n) = (-1)^n$ where $n \in N$ (4)

Following identity connects three neighboring Fibonacci numbers:

$$F_{n-1} + F_n + F_{n+1} = (-1)^n \qquad (5)$$

Also, $Q^n = Q^{n-1} + Q^{n-2}$ (6)

$\Rightarrow$ $Q^{n-2} = Q^n - Q^{n-1}$ (7)

where:

$$Q^{-n} = \begin{bmatrix} F_{n-1} & -F_n \\ -F_n & F_{n+1} \end{bmatrix} \qquad (8)$$

C). Fibonacci encryption/decryption algorithm:

The concept of Fibonacci Q-matrices allows us to develop a symmetric algorithm. This algorithms assumes an initial message in the form of square matrix M of size (p+1) x (p+1) where p = 0, 1, 2, 3,…. Now choose the Fibonacci Q_p-matrix, Q_p^n, of size (p+1) x (p+1) as a encryption (key) matrix and it's inverse matrix, Q_p^{-n}, of the same size as decryption (key) matrix. Therefore, the encryption and decryption are defined by parameters n and p.

Encryption algorithm:

Algorithm encrypt(M)

1. Choose n
2. Choose p
3. Compute Q_p^n
4. $E \leftarrow M \times Q_p^n$ // Compute Cipher text
5. End of algorithm

The working of above symmetric key encryption algorithm based on classical Q-matrix is beautifully illustrated in [8] and [9].

Step 1: Let plain text message is :

$$M = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix}$$

Where $m_i > 0$; i = 1, 2, 3, 4.

Step 2: Choose n = 6 and p = 1 such that

$$Q^6 = \begin{bmatrix} 13 & 8 \\ 8 & 5 \end{bmatrix}$$

Step 3:

$$M \times Q^6 = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \times \begin{bmatrix} 13 & 8 \\ 8 & 5 \end{bmatrix} = \begin{bmatrix} 13m_1 + 8m_2 & 8m_1 + 5m_2 \\ 13m_3 + 8m_4 & 8m_3 + 5m_4 \end{bmatrix}$$

This implies that,

$$c_1 = 13m_1 + 8m_2$$
$$c_2 = 8m_1 + 5m_2$$
$$c_3 = 13m_3 + 8m_4$$
$$c_4 = 8m_3 + 5m_4$$

Decryption algorithm:

Algorithm decrypt (n, p, E)

1. Compute Q_p^{-n}
2. $M \leftarrow E \times Q_p^{-n}$ // Generate Plain text.
3. End of algorithm

A. P. Stakhov et al [8, 9] explained the decryption process as follows:

Step 1: Received encoded message is represented in the matrix form

$$E = \begin{bmatrix} e_1 & e_2 \\ e_3 & e_4 \end{bmatrix}$$

Step 2: Compute the reversible decryption function:

$$Q^{-6} = \begin{bmatrix} 5 & -8 \\ -8 & 13 \end{bmatrix}$$

Step 3: Recover plain text

$$M = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} = \begin{bmatrix} e_1 & e_2 \\ e_3 & e_4 \end{bmatrix} \times \begin{bmatrix} 5 & -8 \\ -8 & 13 \end{bmatrix}$$

Above algorithm has been analyzed, implemented and tested [10] and they concluded that the algorithm works faster than symmetric algorithms (including DES, 3DES, AES and Blowfish).

In this paper, the idea of time varying key (using Fibonacci Q-matrix) has been suggested. Very little work has been done in this direction. P. Chakrabarti et al [7] proposed some approaches which are yet to be investigated experimentally.

III. PROPOSED METHOD

In this section we propose a technique where the key is made to vary from session to session. Hence even if the cryptanalyst gains the key of previous session, it would be theoretically impossible for extracting the original message in next session. We also enhance the security of algorithm based on the *automatic key variability* concept applied on the message based on Fibonacci function. Sender uses Fibonacci Q-matrix, with automatic variable key, to generate cipher text. Receiver can then apply inverse operation of this Fibonacci Q-Matrix to decipher to recover original message [8]. The scope and beauty of this Fibonacci function is that it is a reversible function (similar to XOR). Reversible functions are needed for successful operation of symmetric algorithms. So, the main focus is on the investigation of such reversible function.

Automatic Variable symmetric Key using Fibonacci Q-matrix:

From previous discussion it can be seen that Fibonacci Q-matrix is a powerful technique for securing input data and files of varying content and sizes. And as we have pointed out that state-of-art practices (for increasing the security of information transmission) rely on increasing the key size that consumes time and thus requiring more computation power as well as battery power. We claim here that Fibonacci Q-matrix can be used as reversible function and it can be used for automatic variability of key. The Q matrix at a particular session of given n and p values contains F_{n-1}, F_n and F_{n+1}. Thus for one session the sender and receiver not only have the key of current session but also probable keys of previous and next session. Here key (n, p) is made to vary from session to session hence even if the intruder gets unwanted access to the key of session i, it would not be valid for original message extraction in session i+1 onwards. This enhances the security of algorithm and using the reversibility of Fibonacci Q-Matrix the receiver will receive the data correctly after the application of Q_p^{-n}. Further performance enhancement can be made in the symmetric key exchange over traditional ones that instead of exchanging entire key over communication channel, we pass the parameters only.

IV. FUTURE WORK AND SCOPE.

In this paper we presented a model for investigation of cipher generation technique based on variability concept in Fibonacci Q-matrix. The design of alternative approaches for symmetric key algorithms based on variability of key instead of increasing key size is the biggest challenge, It's vulnerability from intruders point of view may be another direction in this regards.

ACKNOWLEDGMENT

We would like to thank Dr. P. Chakrabarti for their valuable suggestions, support and guidance. This work is supported by research grant from MANIT, Bhopal, India under Grants in Aid Scheme 2010-11, No. Dean(R&C)/2010/63 dated 31/08/2010.

REFERENCE :

[1] S. William and Stalling, Cryptography And Network Security, 4/E. Pearson Education India, 2006.

[2] B. Schneier, Applied cryptography: protocols, algorithms, and source code in C. Wiley, 1996.

[3] Maxime Fernández1, Gloria Diaz1, Alberto Cosme1,Irtalis Negrón1, Priscilla Negrón1, Alfredo "Cryptography: algorithms and security applications" The IEEE Computer Society's Student Fall 2000 Vol. 8 No. 2.

[4] D. S. A. Elminaam, H. M. A. Kader, and M. M. Hadhoud, "Performance Evaluation of Symmetric Encryption Algorithms," International Journal of Computer Science and Network Security, vol. 8, no. 12, pp. 280–286, 2008.

[5] D. S. A. Elminaam, H. M. A. Kader, and M. M. Hadhoud, "Performance Evaluation of Symmetric Encryption Algorithms on Power Consumption for Wireless Devices," International Journal of Computer Theory and Engineering, vol. 1, no. 4, pp. 1793–8201, 2009.

[6] M. Hellman, "An overview of public key cryptography," IEEE Communications Magazine, vol. 16, no. 6, pp. 24–32, 1978.

[7] P. Chakrabarti, B. Bhuyan, A. Chowdhuri, and C. Bhunia, "A novel approach towards realizing optimum data transfer and Automatic Variable Key (AVK) in cryptography," IJCSNS, vol. 8, no. 5, p. 241, 2008.

[8] Stakhov A.P., "Fibonacci matrices, a generalization of the 'Cassini formula', and a new coding theory," Chaos, Solitons & Fractals, vol. 30, no. 1, pp. 56–66, Oct. 2006.

[9] A. NALLI, "On the Hadamard Product of Fibonacci Qn matrix and Fibonacci Q- n matrix," Int. J. Contemp. Math. Sciences, vol. 1, no. 16, pp. 753–761, 2006.

[10] A. Nadeem and M. Y. Javed, "A performance comparison of data encryption algorithms," in Information and Communication Technologies, 2005. ICICT 2005. First International Conference on, 2005, pp. 84–89.

◈◈◈

A Study of the Applications of Data Mining Techniques in Higher Education

[1]Sushil Verma, [2]R. S. Thakur & [3]Shailesh Jalori

[1]Department of Computer Applications, SATI Vidisha
[2]Department of Computer Applications, MANIT, Bhopal (MP), India
[3]Department of applied Maths & Computer science, SATI, Vidisha

Abstract - Data mining is used to extract meaningful information and to develop significant relationships among variables stored in large data set. Few years ago, the information flow in education field was relatively simple and the application of technology was limited. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. Today, one of the biggest challenges that educational institutions face is the explosive growth of educational data and to use this data to improve the quality of managerial decisions and student's performance. The main objective of higher education institutions is to provide quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge 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. The paper aims to purpose the use of Data mining techniques to improve the efficiency of higher educational institutions. If data mining techniques such as clustering, dicision tree and association can be applied to higher education processes, it can help improve student's performance.

Keywords- *Data Mining, Educational Data Mining, Clustering, Decision tree, Classification, Prediction.*

I. INTRODUCTION

The concept of data mining is the technique of extracting previously unknown information with the widest relevance from databases, in order to use it in the decision-making process. Data Mining is a process of extracting previously unknown, valid, potentional useful and hidden patterns from large data sets [1]. As the mount of data stored in Educational database is increasing rapidly. In order to get required benefits from such large data and to find hidden relationships between variables using different data mining techniques developed and used. Clustering and decision tree are most widely used techniques for future prediction. The main goal of clustering is to partition students into homogeneous groups according to their characteristics and abilities [2].Nowadays; higher educational organizations are placing in a very high competitive environment and are aiming to get more competitive advantages over the other business competitors. Data mining techniques are analysis tool that can be used to extract meaningful knowledge from large data sets [3].Educational data mining uses many techniques such as decision trees, neural networks, k-nearest neighbor, naive bayes, support vector machines and many others [4]. Decision tree analysis is a popular data mining technique that can be used to explain the interdependencies among different variables such as attendance ratio and grade ratio. Clustering is one of the basic techniques often used in analyzing data sets. Data mining encompasses different algorithms that are diverse in their methods and aims [5].Today, data collecting and storing are no longer expensive and difficult task. As a result, datasets are growing explosively. To extract the knowledge and information from these massive datasets has attracted a great deal of scientific attention and has become an important research area. Data mining is a flourishing research field and has become a synonym for the process of extracting hidden and useful information from datasets. Data mining provides many tasks that could be used to study the student performance.

II. DATA MINING TECHNIQUES

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. Various algorithms and techniques like Classification, Clustering, Prediction, Association

Rule, Decision Trees, Outlier etc., are used for knowledge discovery from databases. These techniques and methods in data mining need brief mention to have better understanding.

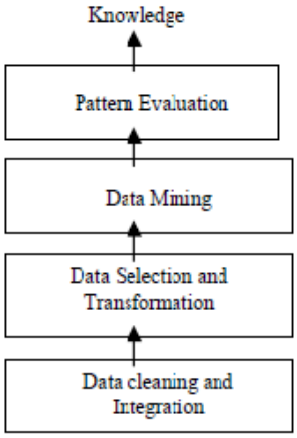


Figure 1: The steps of extracting knowledge from data

A) Clustering: -

Clustering is a technique by which similar records are grouped collectively. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is sometimes used to denote segmentation. Clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with [6]. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is Formed can be viewed as a class of objects, from which rules can be derived. Application of clustering in education can help institutes group individual student into classes of similar behavior, Partition the students into clusters, so that students within a cluster (e.g. Average) are similar to each other while dissimilar to students in other clusters (e.g. Intelligent, Weak).

B) Classification and Prediction: -

Classification is the most commonly applied data mining Technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or Unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as prediction. 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 [7].

C) Association rule: -

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely Used for market basket or transaction data analysis. Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis [3]. Association Rule algorithms need to be able to generate rules with confidence Values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

D) Decision tree:-

A decision tree is a foretelling model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their categorization. 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 and Chi Square Automatic Interaction Detection.

E) Outlier: -

A database may contain data objects that do not comply with the general behavior of the data and are called outliers. The analysis of these outliers may help in fraud detection and predicting abnormal values.

III. RELEATED 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. ("Han j. and Kamber", 2006) Explained that k-means is a well known clustering algorithm tends to uncover relations among variables already presented in dataset [3]. ("Kifaya", 2009) explained that K-means clustering is a widely used method that is easy and quite simple to understand. Cluster analysis describes the similarity between different cases by calculating the distance. These cases are divided into different clusters due to their similarity [5]. ("Henrik", 2001) concluded that clustering was Effective in finding hidden relationships and associations between different categories of students [8]. ("Galit.et.al ", 2007) 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 [9]. ("Erdogan and Timor ", 2005) Used educational data mining to identify and enhance educational process which can improve their decision making process [10]. ("Mohammad Reza Beikzadeh", 2004) in order to analyze student's trends and behaviors toward education. Lack of deep and enough knowledge in higher educational system may prevent system management to achieve quality objectives, Data Mining methodology can help bridging this knowledge gaps in higher education system [11]. ("Maclennan.J", 2005) Data mining techniques can be utilized effectively in selecting course, managing student's improving attendance / dropouts providing supplementary classes where necessary, allocating instructors in a better managed way and thus improving overall stature of the institute/university [12]. ("K. H. Rashan", 2011) A Data Mining model can monitor each student's progress by capturing the variables such as previous semester grade, test mark, assignment grade and attendance. The student's performance can also be analyzed based on the features of interpersonal peer groups such as intellectual self confidence, scoring pattern and time spent with peer groups [13]. ("Luan", 2002) described that higher education institutions carry three duties that are Data Mining intensive [14]. They are:

1. Scientific research that relates to the creation of knowledge.
2. Teaching that concerns with the transmission of knowledge.
3. Institutional research that pertains to the use of knowledge for decision making.

(" Brijesh Kumar Baradwaj, ",2011) described that the hidden patterns, associations and anomalies, which are discovered by some Data mining techniques, can be used to improve the effectiveness, efficiency and the speed of the processes [15].

IV. RESEARCH METHODOLOGY

The research methodology adapted is based on the in-depth study of the topic pertaining to the data mining and its application in higher education. The views of various national and international conferences were taken into consideration while analyzing the data mining applications in the field of higher education. The talks with various academicians, institutions, colleges offering higher education and experts in the field of data mining helped us to find and present the techniques, process and application of data mining in higher education. The main objective of higher education institutions is to provide quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge 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. Higher educational system can be viewed from two different angles. One is the outsider environmental motivation and the other is internal educational reason. The outsider environmental motivation can be observed in higher educational system where the higher educational organizations are aiming to be ahead of their business competitors. Therefore they first have to be powered by a proper roadmap and to be demonstrated with an exact guideline of reaching the top-level educational level. The internal educational reason is considered as the proceeding toward improving the educational management system. Higher educational organization is projected with the promises of more speed during the processes, fewer costs, more quality and flexibility. We have identified six main processes in higher educational systems, which are "evaluation, planning, registration, consulting, marketing and examination". Each process can be categorized into some sub-process. As an example, "evaluation", is an educational process. Its main sub-process are "student assessment", "lecturer assessment", "industrial training assessment", "course assessment" and "student registration evaluation". The main idea in our proposed work is improving the current processes to some new and enhanced educational processes, which have got superior advantages over the traditional processes."Course selection consulting" is a sub-process under the "consulting" educational main process. By applying some of the classification, clustering or association technique on the set of student taken various courses data, the characteristic patterns of previous

students who took particular elective subjects or courses, and the association of courses or elective subject by various type of student can be extracted as knowledge and be stored in knowledgebase. The resulting enhanced process is "classification or association students to the most appropriate course and elective subject". The output of the process can be used by the faculty's consultant to present the most suited courses to students and also by educational course planner to have more advanced strategies on student's course planning.

V. CONCLUSION

Among several innovation in recent technology, data mining is making comprehensive changes in the field of higher education. We have discussed the various data mining techniques which can support education system via generating strategic information. Since the application of data mining brings a lot of advantages in higher learning institution, it is recommended to apply these techniques in the areas like optimization of resources, prediction of retainment of faculties in the university. Data mining techniques capabilities provided effective improving tools for student performance. It showed how useful data mining can be in higher education in particularly to predict the final performance of student.

REFERENCE :

[1] C. Romero, S. Ventura, E. Garcia (2008), "Data mining in course management Systems: Moodle case study and tutorial", Computers & Education, Vol. 51, No. 1, pp. 368-384, 2008.

[2] Connolly T., C. Begg and A. Strachan (1999), "Database Systems: A Practical Approach to Design, Implementation, and Management" (3rd Ed.). Harlow: Addison-Wesley.687, 1999.

[3] Han, J. and Kamber, M., (2006), "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems", Jim Gray, Series Editor, 2006.

[4] Wayne Smith (2005), "Applying Data Mining to Scheduling Courses at a University", Communications of the Association for Information Systems, Vol. 16, Article 23, 2005.

[5] Kifaya (2009),"Mining student evaluation using associative classification and clustering", Communications of the IBIMA vol. 11 IISN 1943-7765, 2009.

[6] C. Romero, S. Ventura (2007), "Educational data mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146, 2007.

[7] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan (2010), "Data Mining Model for Higher Education System", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.

[8] Henrik (2001),"Clustering as a Data Mining Method in a Web-based System for Thoracic Surgery": © 2001.

[9] Galit.et.al (2007), "Examining online learning processes based on log files analysis": a case study, Research, Reflection and Innovations in Integrating ICT in Education, 2007.

[10] Erdogan and Timor (2005),"A data mining application in a student database", Journal of Aeronautic and Space Technologies July 2005 Volume 2 Number 2 (53-57), 2005.

[11] Mohammad Reza Beikzadeh, Naeimeh Delavari (2004),"A New Analysis Model for Data Mining Process in Higher Educational Systems" TS3B-2 in M2USIC 2004.

[12] ZhaoHui. Maclennan.J, (2005)," Data Mining with SQL Server 2005" Wihely Publishing, Inc, 2005.

[13] K. H. Rashan, Anushka Peiris (2011), "Data Mining Applications in the Education Sector", MSIT, Carnegie Mellon University, retrieved on 28/01/2011.

[14] Luan, J. (2002),"Data mining and Its Applications in Higher Education" in A.Serban and J.Luan (eds.) Knowledge Management: Building a Competitive Advantages for Higher Education. New Directions for institutional Research, No.113.San Francisco, CA: Jossey Bass.

[15] Brijesh Kumar Baradwaj, Sourabh Pal (2011),"Mining Educational Data to Analyze Student's Performance", IJACSA Volume 2, no. 6, 2011.

◈◈◈

A Brief Study of Security Issues in Cloud Computing

[1]Urmila Mahor & [2]R. S. Thakur
[1]Bansal Institute of Science & Technology, Bhopal
[2]Moulana Azad National Institute, Bhopal

Abstract - Cloud computing is today's most enticing technology areas due to its cost-efficiency and flexibility. However, despite the surge in activity and interest, there are significant, persistent concerns about cloud computing that are impeding momentum and will eventually compromise the vision of cloud computing as a new IT procurement model

In this paper, I characterize the problems and their impact on adoption. In addition, and equally importance, I describe how the combination of existing research thrusts has the potential to alleviate many of the concerns impeding adoption. In particular with continued research advances in trusted computing and computation-supporting encryption, life in the cloud can be advantageous from a business intelligence standpoint over the isolated alternative that is more common today.

In this paper I tried to explain the cloud computing along with its secure issues and challenges in brief and emphasize on various security threats in cloud computing also the existing methods to control them along with their pros and cons.

General Terms : *Security, Standardization, Legal Aspects.*

Keywords : *Cloud computing, security, privacy*

I. INTRODUCTION

Cloud computing is the collection of virtualized and scalable resources, capable of hosting application and providing required services to the users with the "pay only for use" strategy where the users pay only for the number of service units they consume. A computing Cloud is a set of network enabled services, providing scalable, quality of services guaranteed, normally personalized, inexpensive computing infrastructures on demand, which could be accessed in a simple and pervasive way. Despite of this, advantages such as On demand infrastructure, pay as you go, reduced cost of maintenance, elastic scaling etc. are compelling reasons for enterprises to decide on cloud computing environments.

Usually, in a cloud computing paradigm, data storage and computation are performed in a single datacenter. There can be various security related advantages in using a cloud computing environment. However, a single point of failure can not be assumed for any data loss. As shown in Figure 1, the data may be located at several geographically distributed nodes in the cloud. There may be multiple points where a security breach can occur. Compared to a traditional in house computing, it might be difficult to track the security breach in a cloud computing environment. In this paper, I present the advantages and disadvantages (in the context of data security) of using a cloud environment. I carry out a small survey on major cloud service providers to investigate the prominent security issues.

In order to build a better trust mechanism, I present a risk analysis approach that can be primarily used by the perspective cloud users before putting their confidential data into a cloud my approach is based on the idea of trust model, principally used in distributed information systems [1,2]. I express the general idea of trust management and present its use in analyzing the data security risks in cloud computing.

The concept of this paper is summarized as follows:

(a) Discuss the major security issues in cloud computing paradigms.

(b) A brief a survey of major cloud service providers to explore the security mechanisms in the context of security issues.

(c) A brief idea of a risk analysis approach that can be used by a prospective cloud service user toevaluate the risk of data security.

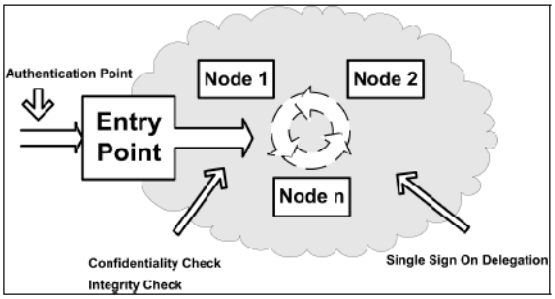


Fig. 1. Typical Data Security Checkpoints in a Cloud Computing Environment

II. SECURITY ISSUES AND CHALLENGES

IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS (Software as a Service) are three general models of cloud computing. Each of these models possess a different impact on application security [3]. However, in a typical scenario where an application is hosted in a cloud, two broad security questions that arises are:

– *How secure is the Data?*

– *How secure is the Code?*

Cloud computing environment is generally assumed as a potential cost saver as well as provider of higher service quality. Security, Availability, and Reliability are the major quality concerns of cloud service users. Gens et. al. [4],suggests that security in one of the prominent challenge among all other quality hallenges.

2.1 Security Threats

Cheap data and data analysis. The rise of cloud computing has created enormous data sets that can be monetized by applications

Cost-effective defense of availability. Availability also needs to be considered in the context of an adversary whose goals are simply to sabotage activities.

Increased authentication demands. The development of cloud computing may, in the extreme, allow the use of thin clients on the client side. Rather than a license purchased and software installation on the client side, users will authenticate in order to be able to use a cloud application.

Mash-up authorization. As adoption of cloud computing grows, we are likely to see more and more services performing mash-ups of data. This development has potential security implications, both in terms of data leaks, and in terms of the number of sources of data.

2.2 Other Security Threats

1. **Failures in Providers Security.** Cloud providers control the hardware and the hypervisors on which data is stored and applications are run and hence their security is very important while designing cloud.
2. **Attacks by other customer.** If the barriers between customers break down, one customer can access another customer's data or interfere with their applications.
3. **Availability and reliability issues**. The cloud is only usable through the Internet so Internet reliability and availability is essential.
4. **Legal and Regulatory issues**. The virtual, international nature of cloud computing raises many legal and regulatory issues regarding the data exported outside the jurisdiction.
5. **Perimeter security model broken.** Many organizations use a perimeter security model with strong security at the perimeter of the enterprise network. The cloud is certainly outside the perimeter of enterprise control but it will now store critical data and applications.
6. **Integrating Provider and Customer Security Systems**

Cloud providers must integrate with existing systems or the bad old days of manual provisioning and uncoordinated response will return.

3. Security in clouds

Categories of the security concerns as:

- Traditional security
- Availability
- Reliability

Traditional security - this category include

1. VM-level attacks. Potential vulnerabilities in the hypervisor or VM technology used by cloud vendors are a potential problem in multi-tenant architectures. Vulnerabilities have appeared in VMWare [7], Xen [8], and Microsoft's Virtual PC and Virtual Server [6]. Vendors such as Third Brigade [5] mitigate potential VM-level vulnerabilities through monitoring and firewalls.

2. Cloud provider vulnerabilities. These could be platform-level, such as an SQL-injection or cross-site scripting vulnerability in salesforce.com. For instance, there have been a couple of recent Google Docs vulnerabilities [9] and [12]. The Google response to one of them is here: [10]. There is nothing new in the nature

of these vulnerabilities; only their setting is novel. In fact, IBM has repositioned its Rational AppScan tool, which scans for vulnerabilities in web services as a cloud security service (see Blue Cloud Initiative [11]).

3. Phishing cloud provider. Phishers and other social engineers have a new attack vector, as the Salesforce phishing incident [13] shows.

4. Expanded network attack surface. The cloud user must protect the infrastructure used to connect and interact with the cloud, a task complicated by the cloud being outside the firewall in many cases. For instance, [14] shows an example of how the cloud might attack the machine connecting to it.

5. Authentication and Authorization. The enterprise authentication and authorization framework does not naturally extend into the cloud. How does a company meld its existing framework to include cloud resources? Furthermore, how does an enterprise merge cloud security data (if even available)with its own security metrics and policies?

6. Forensics in the cloud. "Traditional digital forensic methodologies permit investigators to seize equipment and perform detailed analysis on the media and data recovered. The likelihood therefore,of the data being removed, overwritten, deleted or destroyed by the perpetrator in this case is low.More closely linked to a CC environment would be businesses that own and maintain their own multi-server type infrastructure, though this would be on a far smaller scale in comparison. However,the scale of the cloud and the rate at which data is overwritten is of concern."

Availability

1. Uptime. As with the Traditional Security concerns, cloud providers argue that their server uptime compares well with the availability of the cloud user's own data centers.
2. Single point of failure. Cloud services are thought of as providing more availability, but perhaps not there are more single points of failure and attack.
3. Assurance of computational integrity. Can an enterprise be assured that a cloud provider is faithfully running a hosted application and giving valid results

Reliability

There is also a potential lack of control and transparency when a third party holds the data. Part of the hype of cloud computing is that the cloud can be implementation independent, but in reality regulatory compliance requires transparency into the cloud.

IV. EXISTING SOLUTIONS FOR SECURITY THREATS

4.1 Mirage Image Management System

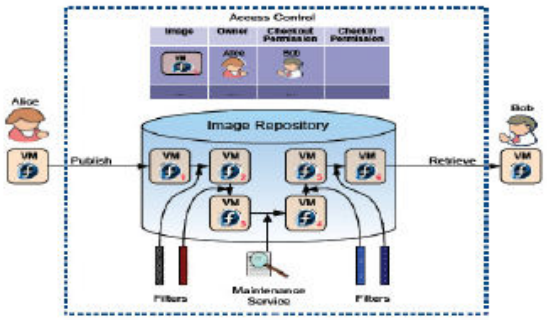


Figure 2 shows the overall architecture of Mirage Image Management System.

Mirage Image Management System consists of 4 major components:

1. **Access Control**. This framework regulates the sharing of VM images. Each image in the repository has a unique owner, who can share images with trusted parties by granting access permissions.

2. **Image Transformation by Running Filters.** Filters remove unwanted information from images at publishes and retrieval time. Filters at publish time can remove or hide sensitive information from the publisher's original image. Filters at retrieval time filters may be specified by the publisher or the retriever.

3. **Provenance Tracking.** This mechanism that tracks the derivation history of an image.

4. **Image maintenance**. Repository maintenance services, such as periodic virus scanning, that detect and fix vulnerabilities discovered after images are published.

Advantages. Filters mitigate the risk in a systematic and efficient way. The system stores all the revisions which allows the user to go back to the previous version if the current version if she desires. The default access permission for an image is private so that only owner and system administrator can access the image and hence untrusted parties cannot access the image.

Limitations. Huge performance overheads, both in space and time. Filters cannot be 100% accurate and hence the system does not eliminate risk entirely. Virus scanning does not guarantee to find all malware in an image. "The ability to monitor or control customer content" might increase the liability of the repository provider (For detailed explanation about Mirage Image Management System please refer [15]).

4.2 Client Based Privacy Manager

Client based privacy manager helps to reduce the risk of data leakage and loss of privacy of the sensitive data processed in the cloud, and provides additional privacy related benefits.

The main features of the privacy manager are:

Obfuscation. This feature can automatically obfuscate some or all of the fields in a data structure before it is sent off to the cloud for processing, and translate the output from the cloud back into de-obfuscated form. The obfuscation and de-obfuscation is done using a key which is chosen by the user and not revealed to cloud service providers.

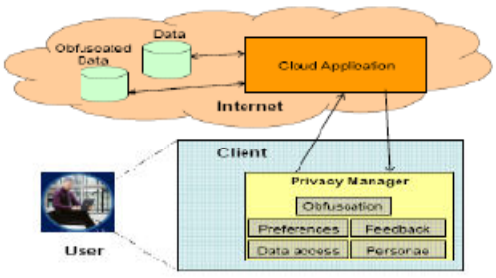


Figure 3 shows the overall architecture of the privacy manager.

Preference Setting. This is a method for allowing users to set their preferences about the handling of personal data that is stored in an unobfuscated form within the cloud. This feature allows the user greater control over the usage of his data.

Data Access. The Privacy Manager contains a module that allows users to access personal information in the cloud, in order to see what is being held about them, and to check its accuracy. This is an auditing mechanism which will detect privacy violations once they have happened.

Feedback. The Feedback module manages and displays feedback to the user regarding usage of his personal information, including notification of data usage in the cloud. This module could monitor personal data that is transferred from the platform.

Personae. This feature allows the user to choose between multiple personae when interacting with cloud services. (Please refer [16] for detailed description of Privacy Manager).

Advantages. This solution solves many practical problems such as Sales Force Automation Problem, Customized End-User Services Problem and Share Portfolio Calculation problem. (Please refer [16] for detailed explanation of these solutions)

Disadvantages. If the service provider does not provide full cooperation the features of the Privacy Manager other than obfuscation will not be effective, since they require the honest cooperation of the service provider. The ability to use obfuscation without any cooperation from the service provider depends not only on the user having sufficient computing resources to carry out the obfuscation and deobfuscation, but also on the application having been implemented in such a way that it will work with obfuscation.

V. CONCLUSION

Cloud computing is the most popular notion in IT today; even an academic report from UC Berkeley says "Cloud Computing is likely to have the same impact on software that foundries have had on the hardware industry." They go on to recommend that "developers would be wise to design their next generation of systems to be deployed into Cloud Computing". my vision also relates to likely problems and abuses arising from a greater reliance on cloud computing, and how to maintain security in the face of such attacks. Namely, the new threats require new constructions to maintain and improve security. Among these are tools to control and understand privacy leaks, perform authentication, and guarantee availability in the face of cloud denial-of-service attacks. In a recent study, a team of computer scientists from the University of California, San Diego and Massachusetts Institute of Technology examined the widely-used Amazon EC2 services. They found that'it is possible to map the internal cloud infrastructure, identify where a particular target VM is likely to reside, and then instantiate new VMs until one is placed co-resident with the target' (Ristenpart et al. 2009: 199). The most obvious finding to emerge from this study is that, there is a need of better trust management. The security analysis and risk analysis approach will help service providers to ensure their customers about the data security. Similarly, the approach can also be used by cloud service users to perform risk analysis before putting their critical data in a security sensitive cloud. At present, there is a lack of structured analysis approaches that can be used for risk analysis in cloud computing environments. The approach suggested in this paper can be a step towards analyzing data security risks.

REFERENCE :

[1] Andert, D., Wakefield, R., Weise, J.: Trust Modeling for Security Architecture Development (2009), http://www.sun.com/blueprints.

[2] Manchala, D.W.: E-Commerce Trust Matrix and Models (2010).

[3] John, H.: Security Guidance for Critical Areas of Focus in Cloud Computing (2009), http://www.cloudsecurityalliance.org/guidance/ Accessed 2 July 2009).

[4] Gens, F.: IT Cloud Services User Survey, part 2: Top Benefits and Challenges (2008).

[5] Third Bridge http://www.thrirdbridge.com.

[6] VirtualPC vulnerability http:// www.microsoft. com/t technet/security/bulletin/ms0ms07-049.mspx.

[7] Warevulnerabilityhttp://securitytracker.com/alerts /2008/Feb/1019493.html.

[8] Waters, B. and Shacham, H. Compact Proofs of Retrievability. In ASIACRYPT. 2008.

[9] Google's response to docs concerns. http://www.pcworld.com/article/2009/03/just-to clarify.html

[10] Google Docs Glitch Exposes Private Files.

[11] http://www.pcworld.com/article/160927/google_ docs_glitche_esposes_private_files.html blue cloud http://www.releses/26642.wss.

[12] Securtity issues with google Doces http://peekay.org/2009/03/26/security-issues-with-google-docs/.

[13] Salesforce.com Warns customers of Phishing Scam.

[14] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, Matei Zaharia. A view of cloud computing. Communications of the ACM , Volume 53 Issue 4, pages 50-58. April 2010.

[15] Shilpashree Srinivasamurthy, David Q. Liu, Survey on Cloud Computing Security – Technical Report. Department of Computer Science, Indiana University Purdue University Fort Wayne July 2010.

[16] Rose, R.: Survey of System Virtualization Techniques(2004), http:// www. robertwrose.com/ vita/ rose- virtualization.pdf.

◈◈◈