National Level Conference on Research Issues in Computer Applications (NCRICA-14), 21-February-2014



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

ISSN (Online): 2279-0055
ISSN (Print): 2279-0047

Organised by:
Dr. N.G.P. Arts and Science College,
Kalapatti Road, Coimbatore– 641048,
Tamilnadu, India.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)
STEM International Scientific Online Media and Publishing House
Head Office: 148, Summit Drive, Byron, Georgia-31008, United States.
Offices Overseas: India, Australia, Germany, Netherlands, Canada.
Website: www.iasir.net, E-mail (s): iasir.journals@iasir.net, iasir.journals@gmail.com, ijetcas@gmail.com

Editorial Board Members of National Level Conference on Research Issues in Computer Applications (NCRICA-14)

Chief Editor
Prof. R. Kousalya
Head of the Department,
Organizing Secretary -NCRICA'14,
Dr. N.G.P Arts and Science College,
Coimbatore 48, Tamil Nadu, India

Editor
Mr. R.Rajeshkanna,
Assistant Professor,
Dr. N.G.P Arts and Science College,
Coimbatore 48, Tamil Nadu, India

Editor
Mrs. K.Gomathy,
Assistant Professor,
Dr. N.G.P Arts and Science College,
Coimbatore 48, Tamil Nadu, India

TABLE OF CONTENTS
(Special Issue 7, Volume 1, 2014)

Special Issue 7, Volume 1

Paper Code	Paper Title	Page No.
IJETCAS 14-001	NS-2 & Glomosim Simulation based Performance Analysis of MANET Routing Protocols G.Kalpana, R.Aarthi	01-05
IJETCAS 14-002	A Study on Kinect Sensor Methods and Applications Anandakrishnan, Dr.S.Santhosh Baboo	06-11
IJETCAS 14-003	LINK MINING: A New Data Mining Challenge Sathish.N, Aravind.R	12-17
IJETCAS 14-004	Comparative Study on Privacy Preserving Through Anonymization based on Clustering Algorithms over Social Networks Asha Yesbhi.C, Murugesakumar.B	18-21
IJETCAS 14-005	Using the Computer Brain Cognitive Computing Bharathi Anbarasan	22-26
IJETCAS 14-006	A Study on Delay Tolerant Networks Santhi.K, Deepa .J	27-29
IJETCAS 14-007	Blocking Records Through Tb Method for Large Data Warehouse Dharani.R, Jayanth.D	30-36
IJETCAS 14-008	Categorizing Information Retrieval Query Using Classification Tree Algorithm Dr. S. Vijayarani, Mr.S.Dhayanand, Ms. N. Nithya	37-42
IJETCAS 14-009	Survey of Various Security attacks and it issues in MANET R.Rajeshkanna, P.Dineshkumar, S.Poorana Senthilkumar	43-52
IJETCAS 14-010	Applications of spatial data mining using GPS Technology Mrs. D.Gandhimathi, M.Ramya Priya	53-56
IJETCAS 14-011	Heterogeneous MANET Using Multiclass Routing and Media Access Control V.Gayathri, A.Vijaya	57-60
IJETCAS 14-012	Dynamic Storage Security in Cloud Computing V. Sridevi, M.Geetha	61-65
IJETCAS 14-013	Survey of Spectrum Sensing Issues and Attacks in Cognitive Radio Networks K.Gomathy,P.Mohanasundhari, R.Rajesh Kanna	66-73
IJETCAS 14-014	Spilt Multipath Routing with Maximally Disjoint Paths in Ad-Hoc Networks Gomathy.K, Hemapriya.K.E	74-81
IJETCAS 14-015	Analysing Risks on Software Testing Gowdhami.D, ArunaDevi.P	82-86
IJETCAS 14-016	Deblurring of Image using Noise Suppressed Steering Kernel Janet Lobo	87-91
IJETCAS 14-017	Performance of Multi Server Authentication and Key Agreement in Network Security Ms. S.Gowthami, Ms.R.Saveetha	92-96
IJETCAS 14-018	Cloud Computing Security from Single to Multi-Clouds M.Bhuvaneshkumar, G. Anitha, Dr. X. Agnes Kalarani, R. Gunasundari	97-101
IJETCAS 14-019	Effective Data Mining using Neural Networks A.Ishwarya, Mrs. P.Laura Juliet	102-104
IJETCAS 14-020	Multimodal Biometric Scheme using Fingerprint & IRIS Fusion V. Sridevi, P.Vidhya	105-109
IJETCAS 14-021	Color Image Segmentation for Fruit Ripeness Detection Jothipriya.I	110-113
IJETCAS 14-022	Subject Based Clustering of Web-Data Using Association Rule Mining Ms M.Kalpana, Dr.V.Arulmozhi	114-116
IJETCAS 14-023	Content-based Recommendation System for Interactive Digital Televisions Mrs. R.Kousalya, Saranya.K	117-119

IJETCAS 14-024	Unprivileged Black-Box Detection of User-Space Key loggers P.K.Mangayarkarasi, S.Narmadha	120-122
IJETCAS 14-025	Secured Networking using Steganography N. Hemalatha, S.Sivaraja	123-127
IJETCAS 14-026	Data Mining - A Novel Tactic Ms.K.Suguna, Ms.S.Deepa Ranjini, Ms.K.Kalaivani	128-131
IJETCAS 14-027	Biometric Finger print reader using Edge Detection Sumaya Nahar E.	132-136
IJETCAS 14-028	A Survey on Dimensionality Reduction Techniques for Document Clustering C.Rathika, K.Shankri	137-142
IJETCAS 14-029	Mobile Ad Hoc Networking using Wireless Mobile T.Sangeetha, P.Ramesh	143-150
IJETCAS 14-030	Optimal Content Downloading in Vehicular Networks C.P Balasubramaniam, T.Priya	151-155
IJETCAS 14-031	An Efficient and Enhanced path sequence algorithm for missing pages in Community Web R.Kousalya, S.Mohanapriya	156-159
IJETCAS 14-032	Analysis of Elliptic Curve Integrated Encryption System in VANET C.Kumuthini, K.Tharani, P.Krishnakumari	160-163
IJETCAS 14-033	Analysis of Various Position-Based Routing Protocols and Greedy Method in VANET C.Kumuthini, P.KrishnaKumari	164-168
IJETCAS 14-034	A New Algorithm for Inferring User Search Goals with Feedback Sessions K.N.Kuvalayajothi, N.Murugan	169-171
IJETCAS 14-035	Resource allocation using Virtual Machines for Cloud Computing Mrs. K.Lavanya, Mrs. S. Rajanandini	172-177
IJETCAS 14-036	An Efficient Method for the Estimation and Extraction of RGB Color Data in a Masked Image using Scatter Plot D. Napoleon, V.Mageshwari, K.Prabavathi	178-181
IJETCAS 14-037	Traitor tracing and secure transmission of files in networks S.Vasanth Kumar, N.Murugan	182-184
IJETCAS 14-038	Data mining and Clustering Pradeep.K, Meena Preethi.B	185-192
IJETCAS 14-039	Web Personalization on Task Based Modeling in Web Usage Mining Dr. A. Saradha, Mrs. T. Mohana Priya	193-199
IJETCAS 14-040	Replication Attacks Detection protocol in Wireless Sensor Networks L.Devi	200-204
IJETCAS 14-041	Energy Efficient of Connectivity and Coverage Preserving Schemes for Surveillance Applications in WSN S.Meena, L.Devi	205-209
IJETCAS 14-042	Lightweight Authentication Protocols for Low-Cost RFID Tags P.Nagarajan, L.Devi	210-217
IJETCAS 14-043	Reliability Based Routing in Wireless Sensor Networks E.Natarajan	218-223
IJETCAS 14-044	Deep Packet Inspection (DPI) for Internet and Surveillance Mrs.S.Thilagavathi, Ms.N.Nithya	224-226
IJETCAS 14-045	A Survey on Steganography Using DCT & DWT Techniques K.Swathi, G.Sathyavathy	227-231
IJETCAS 14-046	Recognizing Pre and Post Surgery Faces Using Multiobjective Evolutionary Firefly Algorithm A.Nirmala, K.Dilipkumar	232-237
IJETCAS 14-047	Data Stream Classification - A Survey Dr. S. Vijayarani, Ms. R. Prasannalakshmi	238-242
IJETCAS 14-048	Cyber security and Advanced web Vulnerability Scanning Mrs. B. Priyadarshini, Mr.R.Prashanth	243-247

IJETCAS 14-049	Manhattan Grid & Random Way Mobility based on Performance Analysis of MANET Routing Protocols G.Kalpana, A.Preethi	248-251
IJETCAS 14-050	Mobile Ad Hoc Networks Applications and Techniques Mrs.S.Rajanandini, Ms.N.Vikasini	252-256
IJETCAS 14-051	Analysis of Security issues and Bandwidth estimation techniques in MANET R.Rajeshkanna, Dr.A.Saradha	257-263
IJETCAS 14-052	Performance Evaluation of Bandwidth in MANET Mrs.B.Ramya	264-270
IJETCAS 14-053	A review study on Artificial Immune System and its Applications B.Ramya , Dr.K.Arthi, Somasundaram	271-275
IJETCAS 14-054	Optimizing Score and Decision Fusion in Palm and Finger Prints Using Pareto Optimization Techniques M.Ramya	276-280
IJETCAS 14-055	Research on Data Mining Using Neural Network Training Mrs.Renuka Devi.S.V	281-286
IJETCAS 14-056	Applications & Trends in Data Mining P.Rizwan Ahmed	287-290
IJETCAS 14-057	Cloud Computing with Related Enabling Technologies P.Rizwan Ahmed	291-297
IJETCAS 14-058	Green Computing “Futures” P.Rizwan Ahmed, A.Aqueel Ahmed	298-302
IJETCAS 14-059	Mobile Commerce and Related Mobile Security Issues P.Rizwan Ahmed	303-305
IJETCAS 14-060	A Study of Mobile IP P.Rizwan Ahmed	306-317
IJETCAS 14-061	Study on Outliers and Hierarchical Clustering Algorithms used for Outlier Detection Sandra Sagaya Mary.D.A., Tamil Selvi.R, Suganya.R	318-321
IJETCAS 14-062	An Optimal Key Management in Wireless Network B Shanmugapriya, S.Saranya	322-325

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

NS-2 & Glomosim Simulation based Performance Analysis of MANET Routing Protocols

[1]G.Kalpana [2]R.Aarthi

[1]Assistant Professor, [2]Research Scholor,
Department of Computer Science,Sri Ramakrishna College of Arts and Science College,Coimbatore.

ABSTRACT

Mobile ad-hoc networks having no fixed infrastructure and without any physical connections of mobile devices such as laptop etc. Due to mobility of nodes, interference, multipath propagation and path loss. MANET has no fixed topology. Hence, MANET has to need dynamic routing protocols for these networks to function properly and delivering the data packets from source node to desired destination. This property of the nodes makes the mobile ad hoc networks unpredictable from the point of view of scalability and topology. In general, routing protocols for MANETs are designed based on the assumption that all participating nodes are fully cooperative. The most popular ones are AODV and DSR. The performance measurements are based on performance metrics such as packet delivery fraction, average end to end delay. This paper analyzes the performance of using execution time by varying different MANET simulation parameters.

Index Terms – AODV, DSR, Simulation parameters

I. INTRODUCTION

Mobile Ad Hoc Network (MANET) is a collection of communication devices or nodes that wish to communicate without any fixed infrastructure and predetermined organization of available links. The nodes in MANET themselves are responsible for dynamically discovering other nodes to communicate. It is a self-configuring network of mobile nodes connected by wire-less links the union of which forms an arbitrary topology. The routing concept basically involves, two activities: firstly, determining optimal routing paths and secondly, transferring the information groups (called packets) through an internetwork [1]. The later concept is called as packet switching [2] which is straight forward, and the path determination could be very complex. The link of two nodes in the ad hoc network can be either unidirectional or bidirectional so the host must know this information about the neighbours. The Hello messages [3] are broadcasted periodically by nodes for the neighbour sensing to check the existence of node in the network. After receiving Hello message the destination node generate an acknowledgment and send as a reply to source node. The Hello messages are only broadcasted one hop away so that they are not forwarded further. DSR (Dynamic Source Routing) is on-demand, simple and efficient routing protocol for multi-hop wireless ad hoc networks of mobile nodes. DSR uses source routing and the protocol consist of two main mechanisms-'Route Discovery' and 'Route Maintenance', which works together entirely, on demand. The protocol allows multiple routes to destination, loop-free routing, support for unidirectional links, use of only 'soft state' in routing, rapid discovery when routes in the network change, designed for mobile ad hoc networks of about two hundred nodes and to work well even with high rates of mobility. AODV (Ad hoc on-demand distance vector) enables dynamic, self-starting, multi-hop on-demand routing for mobile wireless ad hoc networks. AODV discovers paths without source routing and maintains table instead of route cache. It is loop free using destination sequence numbers and mobile nodes to respond to link breakages, changes in network topology in a timely manner.

The network simulator ns-2 is a commonly used tool for MANETS protocol evaluation. In this study, ns-2.35 is compared with glomosim simulator and the result is produced.

This paper deals with the II.Classification of routing protocol III. Overview of AODV and DSR IV. Simulation based Analysis.

II. CLASSIFICATION OF ROUTING PROTOCOL

A routing protocol [4] is desirable whenever a packet needs to be transmitted to a destination via, number of nodes and many routing protocols have been planned for such kind of ad hoc networks.
MANET routing protocols can be categorized into:
- Proactive routing protocols
- Reactive routing protocols.

A. PRO-ACTIVE OR TABLE DRIVEN ROUTING PROTOCOLS

Proactive MANET protocols are based on periodic exchange of control message and maintain routing table and dynamically determine the layout of the network. Through a regular exchange of network topology packets between the nodes of the network, a complete path of the network is maintained at every single node. Hence minimal delay in determines the route to be taken. Some Proactive MANET protocols include: WRP , DSDV, DBF, GSR.

B. REACTIVE OR ON DEMAND ROUTING PROTOCOLS

In Reactive protocol, a route is discovered only when it is necessary. The protocol tries to discover a route only on demand, when it is necessary. These protocols generate much less control traffic at the cost of latency. Some Reactive MANET Protocols include DSR, AODV[5]

III OVERVIEW OF AODV AND DSR

A. The Ad hoc On Demand Distance Vector (AODV)

The Ad hoc On Demand Distance Vector(AODV) routing algorithm is a routing protocol designed for ad hoc mobile networks [6] [7]. AODV is capable of both unicast and multicast routing [8]. It is an on demand algorithm, meaning that it builds routes between nodes only as desired by source nodes. It maintains these routes as long as they are needed by the sources. Additionally, AODV forms trees which connect multicast group members. The trees are composed of the group members and the nodes needed to connect the members. AODV uses sequence numbers to ensure the freshness of routes. It is loop-free, self-starting, and scales to large numbers of mobile nodes. AODV builds routes using a route request/route reply query cycle. When a source node desires a route to a destination for which it does not already have a route, it broadcasts a route request (RREQ) packet across the network. Nodes receiving this packet update their information for the source node and set up backwards pointers to the source node in the route tables. In addition to the source node's IP address, current sequence number, and broadcast ID, the RREQ also contains the most recent sequence number for the destination of which the source node is aware. A node receiving the RREQ may send a route reply (RREP) if it is either the destination or if it has a route to the destination with corresponding sequence number greater than or equal to that contained in the RREQ. If this is the case, it unicasts a RREP back to the source. Otherwise, it rebroadcasts the RREQ. Nodes keep track of the RREQ's source IP address and broadcast ID. If they receive a RREQ which they have already processed, they discard the RREQ and do not forward it. As the RREP propagates back to the source, nodes set up forward pointers to the destination. Once the source node receives the RREP, it may begin to forward data packets to the destination. If the source later receives a RREP containing a greater sequence number or contains the same sequence number with a smaller hop count, it may update its routing information for that destination and begin using the better route.

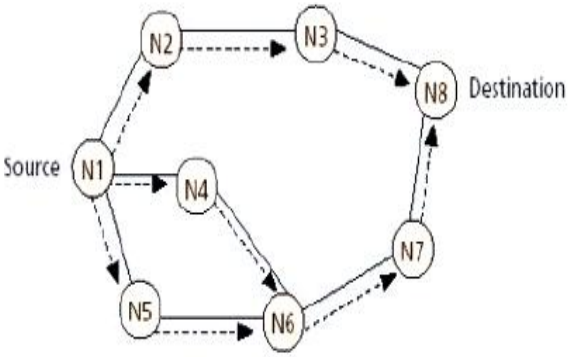


Fig: 1 Route Request (RREQ) flooding

B. Dynamic Source Routing (DSR)

The key distinguishing feature of DSR [4] is the use of source routing. That is, the sender knows the complete hop - by-hop route to the destination. These routes are stored in a route cache. The data packets carry the source route in the packet header. When a node in the ad hoc network attempts to send a data packet to a destination for which it does not already know the route, it uses a route discovery process to dynamically determine such a route. Route discovery works by flooding the network with route request (RREQ) packets. Each node receiving an RREQ rebroadcasts it, unless it is the destination or it has a route to the destination in its route cache. Such a node replies to the RREQ with a route reply (RREP) packet that is routed back to the original source. RREQ and RREP packets are also source routed. The RREQ builds up the path traversed across the network. The RREP routes itself back to the source by traversing this path backward. The route carried back by the RREP packet is cached at the source for future use. If any link on a source route is broken, the source node is notified using a route error (RERR) packet. The source removes any route using this link from its cache. A new route discovery process must be initiated by the source if this route is still needed. DSR makes very aggressive use of source routing and route caching.

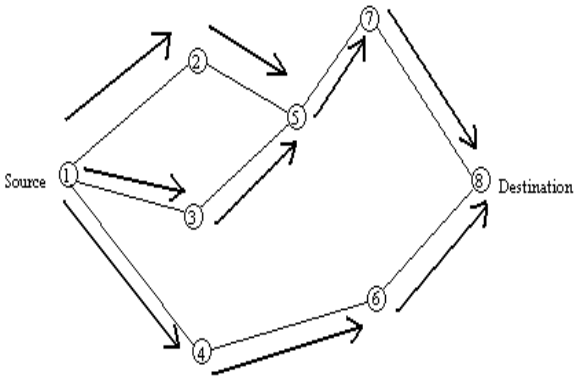


Fig: 2 Propogation of Route Request packet

IV. SIMULATION BASED ANALYSIS

A. Simulation Environment

This section describes the simulation tool, network topology, simulation parameters and simulation result. The performances of AODV, DSR routing protocols are evaluated on the basis of performance metrics given below.

B. Simulation Tools

In this paper the simulation of AODV, DSR routing protocols is done by using network Simulator (NS-2) software due to simplicity and availability. NS is a discrete event simulator targeted at networking research NS provides substantial support for simulation of TCP, routing and multicast routing protocols over wired and wireless network simulator-2 (ver-2.31) is used in the evaluation. The NS instructions can be used to define the topology structure of the network and the motion mode of the nodes, to configure the service source and the receiver, to create the statistical data track file and so on.

All simulations have been carried out using the Glomosim 2.03[1] simulator program under Windows platform. Glomosim is a scalable simulation environment for wired and wireless network systems built using the PARSEC simulation environment. Currently it only supports protocols for a purely wireless network. The modules have been developed using VC++ programming language. It is also built in a layered approach; such as OSI layer network architecture. Glomosim is designed as a set of library modules, each of which simulates a specific wireless communication protocol in the protocol stack. The library has been developed using PARSEC, a C-based parallel simulation language. New protocols and modules can be programmed and added to the library using this language. Glomosim source and binary code can be downloaded only by academic institutions for research purposes.

C.SIMULATION PARAMETERS

Table 1: Simulation Parameter

PARAMETER	***NS-2***	***GLOMOSIM***
Protocol	AODV,DSR	AODV,DSR
No. of nodes	50	50
MAC protocol	MAC/802.11	IEEE802.11
Routing protocol	AODV,DSR	AODV,DSR
Packet size	512 bytes	512 tes

D. PERFORMANCE METRICS

The following performance metrics are used in this paper for performance evaluation of AODV, DSR routing protocols based on NS-2 and Glomosim simulator.

1) *Throughput:* It is the amount of data transferred over the period of time expressed in bits per second or bytes per second.
2) *Packet delivery ratio:* It is the ratio of the number of data packets received by the destination node to the number of data packets sent by the source mobile node.

E. SIMULATION RESULTS

For the glomosim simulation, we will take constant bit rate (CBR) traffic into consideration. In case of both NS-2 and glomosim simulator a comparative graph illustrates the performance and overheads.

THROUGHPUT

Throughput in glomosim network 35 nodes randomly placed with 0s pause time. The results show that for reactive protocols AODV and Proactive protocol WRP, the delivery ratio degrades and raise and then again degrades with increase corresponding to load ratio.

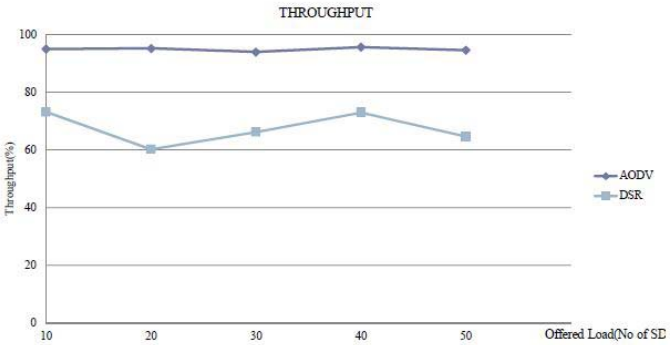


Fig:3 Throughput in Glomosim

From the simulation results in figure 4, it is observed that throughput decreases linearly with increasing number of nodes for both AODV and DSR.

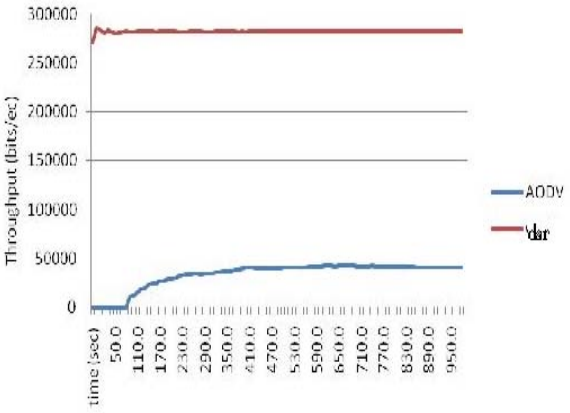


Fig4: Throughput in NS-2

PACKET DELIVERY RATIO

The Fig.5 shows the impact of changing the speed, with which nodes move in an ad hoc network, on the packet delivery ratio. Packet delivery ratio decreases with increase in average node speed in AODV, While packet delivery ratio increase in average node speed in WRP protocol. The packet delivery ratio for AODV is approximately 90%which remains almost same for all node speed. The DSR shows a increase of 30% in delivery ratio when the average node speed increases from 10 m/s to 50 m/s.

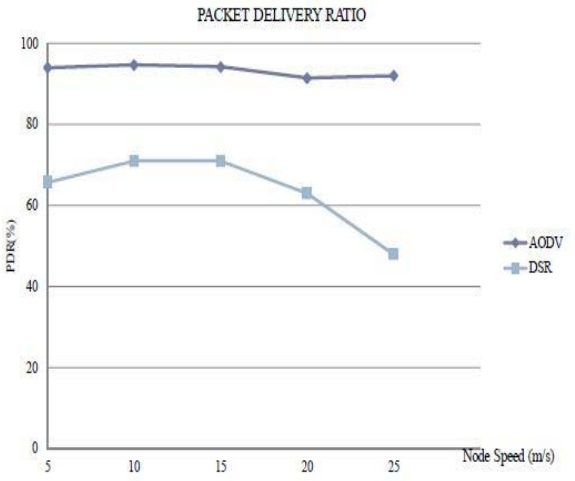


Fig: 5 packet delivery ratio in glomosim

Packet delivery ratio comparison graph shown in fig:6 in case of NS-2 simulator. We observe that DSR shows the lower packet delivery as compared to AODV. So the performance of DSR protocol is better than the AODV routing protocol.

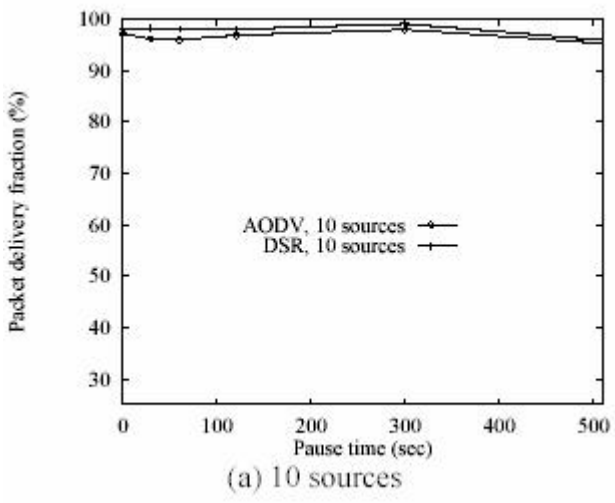


Fig: 6 packet delivery ratio in NS-2

V. CONCLUSION

This paper provides explanation and simulation analysis of on demand routing protocols like AODV and DSR for ad-hoc mobile networks and also provides a classification of these protocols according to the routing strategy (i.e. table driven, on-demand and hybrid routing protocol) on NS-2 and glomosim simulation. Each simulation provides a result and has presented a comparison of these on-demand routing protocol under variation of number of nodes, simultaneously
measured their performances under throughput and packet delivery ratio. From different analysis of graphs and simulations it can be concluded that DSR performs well than that of AODV. DSR protocol shows a best results in and throughput and delivery ratio than AODV protocol. From analysis we observed that packet loss is very less in case of AODV in comparison to DSR. So, we can conclude that if the MANET has to be setup for a small amount of time than AODV.

VI. FUTURE SCOPE

Hybrid routing protocol can be evaluated and tested under NS2 and glomosim based simulator analysis on the same scenario and the result can be determined.

REFERENCES

[1] NS-2, The ns Manual (formally known as NS Documentation) available at http: //www. isi.edu/nsnam/ ns/doc.

[2] R. Al-Ani, "Simulation and performance analysisevaluation for variant MANAT routing protocols",*International Journal of Advancements in Computing Technology, Volume 3,* Number 1, February 2011.-

[3] David Oliver Jorg, : Performance Comparison of MANET Routing Protocols In Different Network Size,Computer Science Project.

[4] Azizol Abdullah, Norlida Ramly, Abdullah Muhammed,Mohd Noor Derahman: Performance Comparison Study ofRouting Protocols for Mobile Grid Environment, pp 82-88,IJCSNS International Journal of Computer Science andNetwork Security, VOL.8 No.2, February 2008

[5]Altman E, and Jimenez T., (2003). NS Simulator forBeginners. Lecture notes. Univ.de Los Andes, Merida, Venezuela and ESSI.Sophia-Antipolis, France.

[6]Charles E. Perkins and Elizabeth M. Royer, "Ad Hoc On Demand Distance Vector Routing," in Proceedings of the2nd IEEE Workshop on Mobile Computing Syslems and Applications, Feb. 1999, pp. 90-100.

[7] B.S. Bakshi, P. Krishna, N.H. Vaidya & D.K. Pradhan, Improving performance of TCP over wireless networks, *Proceeding of the 17th International Conference onDistributed Computing System,* 1997 Page(s): 365-373.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A Study on Kinect Sensor Methods and Applications

[1].Anandakrishnan [2] Dr.S.Santhosh Baboo

[1]Providence College for Women, Coonoor.

[2]PG & Research Dept. of Computer Applications
D.G. Vaishnav College, Chennai.

ABSTRACT

Kinect is a technology that is capable of motion tracking and sensing. Kinect plays a unique role in the study of identification. Its unique properties of being able to identify the objects are performed with a sensor. Kinect Sensors track objects based on the color and data. Kinect sensors have been developed overtime for tracking various actions and postures. Improvisations are inhibited within the technology based on identity, digitization, alpha channels, depth of color and sensors in Autism research. Major concepts in Kinect and their enhancements are surveyed in this paper.

Keywords*:* Kinect, sensor, tracking, color data, depth data

1. Introduction

Multimedia is a field in computer technology that incorporates computer controlled integration of almost all the media. Information is processed and transmitted digitally. Technological advances have lead to the use of these media in real time at a very large amount.

Human computer interaction is becoming highly effective and a requirement. Kinect plays an important role in the communication between human and computer. It uses the study of human language understanding.

Human understanding is all what is required and brought in to the interaction. Understanding the language, the moves, signs and tracking each action or media communicated by the human body can be done with the help of a Kinect. Understanding is a question of "what the user or the human is doing". It is necessary that the system understands what is asked or what is required before the answer is formulated and provided.

Kinect has been developed to overcome the harsh problems faced by video cameras for motion tracking and the process of interaction.

Kinect makes use of a sensor in order to track objects or human beings. These sensors provide a three dimensional depth, interpreting all the moves and actions of the human being. Kinect saw its application and evolution in the gaming industry as shown in *Figure1*. Due to its wide range of efficiency and its practicability, Kinect made its appearance in a number of fields like research, engineering, robotics and others [12].



Figure1. Gaming using Kinect

The impact created by Kinect has spread over a large area covering different fields apart from the gaming industry. Due to its availability and its efficiency it has covered the interest of researchers from different fields including Computer Science, electronics and robotic sensing technology etc. Its applications spread over a wide range from autism research to technology enhancements. Kinect Sensor tracking may depend on signs, gestures, moves or the tracking of the part of the human body like face, hand or the entire body. In order to make the best use of Kinect, it has been improvised and enhanced by using a number of techniques including enhancement in identification, digitization, color and depth data

recognition, segmentation, calibration, research and many others.

2. *Kinect Sensor*

Kinect builds on software technology that introduced a system which can interpret specific gestures. Kinect incorporates a system that is hand free control of the device using a projector, camera and a microchip.

A depth sensor uses a monochrome complementary metal oxide semiconductor sensor. A dot pattern is formed by deducing the relative geometry between the projector and the camera. The values obtained from the Kinect sensor are at times inaccurate. A recalibration technique is used to obtain the affine transformation of depth values with color data as shown in *Figure2*.

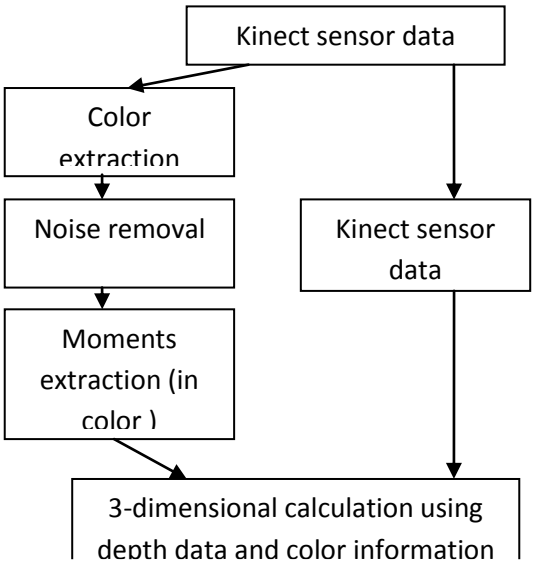


Figure2. Kinect sensing using color and depth data

Because of the calibration produced the depth values produced by the Kinect sensor are inaccurate at times. The recalibration tool providing the affine transformation is used to overcome this problem. With the help of an RGB camera the recalibration is given by 3D coordinates that are true values. The Kinect estimates the values and the transformation between cameras by minimizing the distance between two point classes.

3. *Tracking*

The innovation in hand is about tracking the basic skeleton of objects or the human being. Tracking is influenced by the sought of object and the number. It is also influenced by the environment and the surrounding atmosphere within which tracking can be performed. Bases of tracking are indulged in the parameters such as the size, shape, color, motion and the scares effects produced example furry objects.

Kinect considers the combination of all the parts in an object to identify the object. In a human body, tracking is the detection of head, neck, shoulders, arms etc by their corresponding coordinates thus, executing the entire skeletal tracking mechanism.

Kinect tracking are enhanced and varied across applications.

4. *Some tracking methods*

4.1 Recognition of Sign Languages

Sign Language is based on detecting and translating it to different media. Tracking the signs produced by a human being examines a huge challenge for understanding. Signs are recognized by learning a set of example signs as shown in *Figure3* and discriminating them. The differences and similarities among the signs are studied and classifiers are used to formulate the most discriminative classes. Heuristic distance measures enforce the understanding of the depth data.

Sign Language is an extraordinary communication language for deaf and dumb people. The sign language is a composition of hand and arm movements including facial expressions. Recognition of sign language leads to transferring it to specific text and other media. The two major components in sign language are about extracting information and about how to model the signs and study the similarities and differences among their appearances.

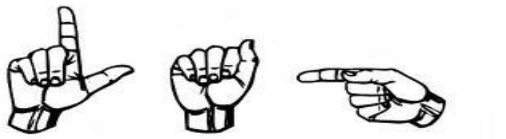


Figure3. Sign Language discrimination

All the frames in the videos combined together given by the sensor consists of a huge amount of differences [15]. Along with all the differences that exits there are other key shapes and similarities with which they are predicted. The key concepts are used as the representative keys for the algorithms used for the recognition of Sign Language.

4.2 Hand Gestures

Human computer interaction [1] involves communication with different media between computer and human. Hand gesture recognition is one amongst it. The performance of the hand gesture recognition is highly affected by the environment, the background and the color depths all of which affect the image in equal proportions as shown in *Figure4*. One way to avoid this is by the use of a different sensor which is not affected by the lighting conditions and color [10]. Now the uses of depth camera have seen a large improvement over the other methods. Because of the low-resolution of

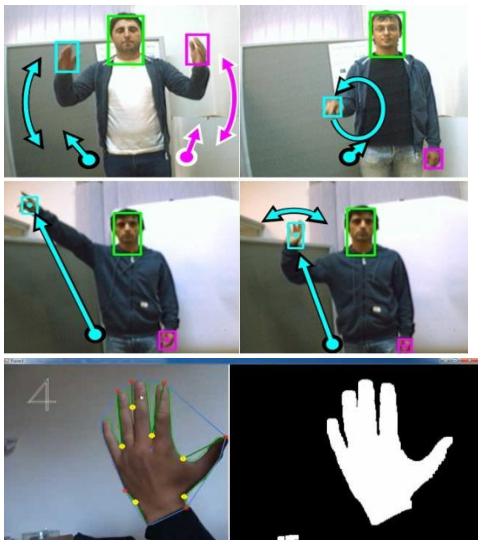


Figure4. Hand Gesture Recognition

Kinect depth map the ability to recognize the hand gesture is very low when compared to the large objects. Kinect sensors involve a huge amount of progress in tracking the objects within the range defined but kinect is seen to produce limitations in robust hand recognition due to their size. Segmentation of these reduced size gestures is a difficult process. Dissimilarities are calculated using a distance measure. In order to detect the noisy hand gestures a new metric called the Finger-Earth Mover's Distance is used to measure the dissimilarities in the hand gestures. The measure is calculated between the smallest of differences in the gesture.

4.3 Face Recognition

A low resolution 3D sensor tracks the face under different conditions as shown in *Figure5*. The pre-processing algorithm obtains the canonical view and the shape of the face. The algorithm used here is used to smooth the noisy depth data given by the sensor. The depth map and the textures are studied and the sparse approximated from training data. The obtained texture is converted to a discriminate color before sparse coding.

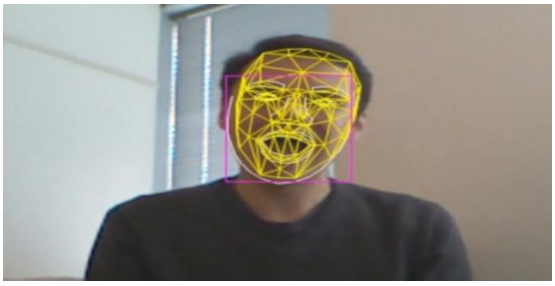


Figure5. Face recognition using 3D sensor

Facial expressions are tracked in order to apply for animation and for the purpose of human computer interaction. They are also influenced by the background and lightning. Applications of this facial recognition using 3D facial scanning for the purpose of morphing produce high quality results.

4.4 Full Human Body Recognition

The entire human body is scanned using multiple Kinects as shown in the *Figure6*. They are position such that they capture the entire human body without any interference and overlapping. Different parts are captured by using a rough mesh template to deform the frames followed by a global alignment in order to distribute the errors, handling closure problems and misalignment [8].

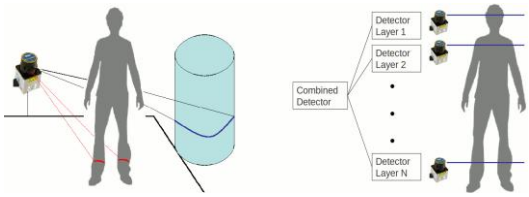


Figure6. Human body recognition using multiple sensors

When a full body scanning is used a rough template of the results to be obtained from the recognition is specified and the scan results produced from the sensor are compared with the results of the template. The alignment process is an involvement of the occlusion process.

5. Enhancement and Improvisation in Kinect

5.1 Kinect Identity

Kinect identity is based on two different techniques using the biometric sign in and the session play. In the sign in technique the user or the human is recognized every time his appearance is made. In the session technique the user's appearance is recognized for a particular time limit beyond which it can be changed. This concept is mainly used in gaming [11].

The observed technique consists of the recognition of the face, color of the clothes and the estimation of height. These characteristics are given for each individual user and if it is a session based concept, the individual characteristics are made valid only for that particular session beyond which his characteristics are invalid and another user session begin. The face micro structure plays an important role in the recognition.

5.2 Object Digitalization

Kinect provides the capability to identify or recognize the 3D objects within the range of the Kinect. The object digitization method allows the digitization of a real time model created to a 3D model inorder to add them as characters within it, solving the very problem of reconstructing 3D models based on the inputs. The techniques use laser lights which are focused on the object to measure the overall surface area and the orientation [13]. The measurement is based to the time taken for the retransmission of the light from the object. In another technique the models created in real life are measured and digitized to form the virtual version of the real life model. This method is an option to people who consider the traditional methods ineffective or costly for the purpose of recognizing the objects.

5.3 Color and Depth Data

The use of both the color and the depth information is a major requirement for scene reconstruction and rendering [5]. In order to bring both color and depth data together there is a requirement for calibrating the cameras separately as well as initializing the calibration among the two. The disparity distortion correction applied involves more accuracy ranges.

Scene reconstruction is a result of the calibration of the system involving the relative calibration of the cameras and the calibration of each [4]. Independent calibration of the system does not produce optimal results whereas the combined calibration does create the required output because of the use of all the information. Certain cameras produce depth and intensity at the same time which simplifies the calibration required. But if device is not known then the image with original intensity cannot be obtained.

5.4 Alpha Channels in Depth Maps

Depth estimation can be done using active or passive sensing methods. The disparity is calculated from the two images obtained in the active and passive methods. But the estimation of the 3D information is a tedious process [7]. In certain techniques the detection of the scene is high but the presence of noise and the low resolution results produced lower the effectiveness of the technique. In others the saturated pixels are not detected. Passive and active sensing separately, have their side effects but these limitations are overcome by using both the sensing techniques together forming a single alpha channel.

At present the depth maps consists of the estimation of the single alpha channel including the advantages of both the passive and active techniques. The Alpha channel method is used in order to overcome the degradation seen in sensing. The depth map is created by the combination of color and the alpha images thus generating outputs with high clarity and naturalness.

5.5 Autism research

The autism individuals unable to communicate with others have increased over time and the need to develop special assessment tools to helps us understand them by engage them to use Kinect based systems. These researches lead to the development of the avatar which is embedded in virtual scenes involving both an individual and a computer model [2].

There is a very less amount of improvement seen in helping such individuals because of the need for assessment tools. If there is a possibility of an individual examiner to attempt to absorb the problems then an understanding can be created. But this is not the case in all the circumstances thus the entertainment technology is used to enhance and develop various assessment tools for understanding, documenting and more realistically communicating.

5.6 Brain-Computer Interface

The system involves the use of the real time driven brain computer interface assisting in the control of an action to be performed [3]. The intention of an action to be performed by the

individual is decoded with the Kinect based vision system.

A common impairment patients are affected by is the stroke. In this case the patient is supported with a rehabilitation robot. Thus avoid the limitations created by the human support like the requirement of the presence of the person over time, the requirement of the actions to be repeated etc.

The robot provides a high intensity motor for task oriented movements of the part of the human body under the specific conditions. The development of this technology was for the very purpose of motivating an affected individual to perform movements by initiating it by him. Impedance assistance, BCI based on electroencephalographic with reality bases or robot assistance, EMG, direct feed back with simulation, etc has been the versions created for the improvement of the brain computer interface.

5.7 Scene Segmentation

Scene Segmentation is a process of identifying different parts in a scene. Scene are characterized by the concept of images. Studying these images is not an easy task. The development of segmentation has seen various methods such as the graph theory, clustering, region matching and many others [9].

The major limitation seen is that the image given is not enough to represent an entire scene. To overcome the complexity of scene segmentation the geometric data is used.

Geometric data along with other details of the image and the scene in combination are taken into consideration. Multidimensional vectors are collectively used to represent both the color and the depth data and the segmentation of the scenes are performed using spectral clustering.

Now tools that support both structured and unstructured scene segmentation are developed. Tools include both the geometric and the color information to obtain a fused framework. The geometric data provided contributes a lot of knowledge about partitioning and about the feature present in the image that forms the scene.

5.8 *Data Compression*

The traditional methods developed in the sensor concepts were based on the effective use of the sensors and the recognition and capturing of the scenes being performed [5][6]. The scene were digitized and used for other purposes such as matching and merging.

The need for transmitting the digitized data and the need to transfer a large amount of data over a particular time lead to the study of data compression. There were a number of changes in the camera to obtain the compression required but due to bulky size and the cost this was unachievable.

Kinect facilitated this possibility to meet the requirement. The depth related data that was considered bulky in the depth sensors given as related applications for easy transmission over the network.

Depth data in Kinect is of a large data range. Depth data are instable thus obviously inhibiting the problem of compression. In order to compresses the data a bilateral filter is used to suppress the noisy data and padding is performed to avoid the invalid data in calibration with coding. Once the data reached the other end it was reconstructed.

1.1 Spinal Loading

Representing the problem of low back disorder there is a use of trunk kinematics for the purpose of supporting biomechanics. The methods used required instrumentation, thus leading to the use of videos, tracking using a single camera etc but they had their limitation of requiring manual operations in joint detection and only a few frames were developed for each task, affecting the acceleration and finally the spinal loading [16]. The success of the human skeleton extraction using the depth sensors lead to the development of tracking the individual objects based on the Spinal loading. Kinect was used in the biomechanical process of detecting the joints and it proved effective and efficient. Kinect is now used for the assessment of the spinal loading. The joints in the skeletal system are predicted and taken advantage of. The movements are captured as parts of a rigid body. The measure of linear and angular acceleration for each part is estimated using kinematics in Kinect.

1.2 Tele collaboration

Most of the collaboration tools example video conferencing etc were of a symmetric view point. The participants of the conference usually have a face to face communication, making the process robust. The symmetric view seen in a number of collaboration tools have proved ineffective [17].

The study on this concept indicated that the limitation was due to the lack of representing

the entire 3D space. One such method is the VR but the system required instrumentation. Another method was using the beaming project that could be used in any kind of environment. But they were influenced by limitations like misuse, transformations etc. Now the use of depth cameras and other assessment tools of kinect have lead to the Asymmetric tool is that the telepresence system that allows the recognition of all the objects within the region as a 3D model enhancing the capabilities of teleconferencing. It is to be enhanced in order to support the multi model representation of the source and destination.

1.3 Out-of-Box Experience

Kinect applications spread over a wide range thus to visualize the use of the product a digital out-of-box experience tool was developed for the Kinect that served as a huge improvisation as for a number of digital devices. OOBE provided the steps and details for the usage of kinect in various applications [14].

2. Conclusion

Kinect plays an important role of recognition and tracking objects. The depth and color data sensors combine together to provide the best scene segmentation required for detection. Digitization of the object and the interfaces created incorporate the possibility of high range of human computer interaction. This paper is a brief study of Kinect, its applications and enhancements for understanding the power and use of Kinect today.

Reference

1) Digan Um, Dongseok Ryu, and Myungjoon Kal, “Multiple Intensity Differentiation for 3D surface reconstruction with Mono-Vision Infrared Proximity Array Sensor” IEEE Sensor Journal, Vol. 11, No. 12, December 2011 1530-437X.

2) Jeff Munson and Phillip Pasqual, “Using Technology in autism Research: The Promise and the Perils” IEEE Computer Society, June 2012, 0018-9162.

3) Antonio Frisoli, Claudio Loconsole, Daniele Leonardis, Filippo Banno, Michele Barsotti, Carmelo Chisari and Massimo Bergamasco, “A New Gaze-BCI-Driven Control of an Upper Limb Exoskeleton for Rehabilitation in Real-World Task” IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, Vol. 42, No.6, November 2012, 1094-6977.

4) Danial Herrera C., Juho Kannala, and Janne Heikkila, “Joint Depth and Color Camera Calibration with Distortion Correction” IEEE transactions and Pattern Analysis and Machine Intelligence, Vol.34, No.10, October 2012, 0162-8828.

5) Dimitrios S. Alexiadis, Dimitrios Zarpalas, and Petros Daras, “Real Time, Full 3D Reconstruction of Moving Foreground Objects from Multiple Consumer Depth Cameras” IEEE Transaction on Multimedia, Vol.15, No.2, February 2013, 1520-9210.

6) Jingjing Fu, Dan Miao, weiren Yu, Shiqi Wang, Yan Lu, and Shipeng Li, “Kinect-Like Depth Data Compression” IEEE Transaction on Multimedia, Vol.15, No.6, October 2013, 1520-9210.

7) Ji-Ho Cho, Kwan H. Lee, and Kiyoharu Aizawa, “Enhancement of Depth Maps with Alpha Channel Estimation for 3D Video” IEEE Journal of Selected Topics in Signal Processing, Vol.6, No.5, September 2012, 1932-4553.

8) Jing Tong, Jin Zhonu, ligang Liu, Zhigeng pan and Hao Yan, “Scanning 3D Full Human Bodies Using Kinects” IEEE Transactions on Visualization and Computer Grapics, Vol.18, No.4, April 2012, 1077-2626.

9) Carlo Dal Mutto, Pietro Zanuttigh, and Guido M.Cortelazzo, “Fusion of Geometry and Color Information For Scene Segmentation” IEEE Journal of Selected Topics in Signal Processing, Vol.6, No.5, September 2012, 1932-4553.

10) Zhou Ren, Junsong Yuan, Jingjing Meng and Zhengyou Zhnag, “Robust Part- Based Hand Gesture Recognition Using Kinect Sensor” IEEE Transactions on Multimedia, Vol.15, No.5, August 2013, 1520-910.

11) Tommer Leyvand, casey Meekhof, Yi-Chen Wei, Jian Sun, and Baining Guo, “Kinect Identity: Technology and Experiemce” IEEE Computer Society, 2011, 0018-9162.

12) Wenjun Zeng, “Microsoft Kinect Sensor and its Effect” IEEE Computer Society, 2012, 1070-986X.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

LINK MINING: A New Data Mining Challenge

[1]Sathish.N [2]Aravind.R

Department of Computer Science,Sri Krishna Arts and Science College,
Coimbatore

ABSTRACT

A key challenge for data mining is tackling the problem of mining richly structured datasets, where the objects are linked in some way. Links among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models. Recently there has been a surge of interest in this area, fuelled largely by interest in web and hypertext mining, but also by interest in mining social networks, security and law enforcement data, bibliographic citations and epidemiological records.

INTRODUCTION

Traditional data mining tasks such as association rule mining, market basket analysis and cluster analysis commonly attempt to find patterns in a dataset characterized by a collection of independent instances of a single relation. This is consistent with the classical statistical inference problem of trying to identify a model given a random sample from a common underlying distribution.

A key challenge for data mining is tackling the problem of mining richly structured, heterogeneous datasets. These datasets are typically multi-relational; they may be described by a relational database, a semi-structured representations such as XML, or using relational or first-order logic. However, the key commonalities are that the domain consists of a variety of object types and objects can be linked in some manner. In this case, the instances in our dataset are linked in some way, either by an explicit link, such as a URL, or by a constructed link, such as a join operation between tables stored in a database. Naively applying traditional statistical inference procedures, which assume that instances are independent, can lead to inappropriate conclusions [24]. Care must be taken that potential correlations due to links are handled appropriately. In fact, record linkage is knowledge that should be exploited. Clearly, this is information that can be used to improve the predictive accuracy of the learned models: attributes of linked objects are often correlated and links are more likely to exist between objects that have some commonality.

Link mining is a newly emerging research area that is at the intersection of the work in link analysis [25; 14], hypertext and web mining [3], relational learning and inductive logic programming [13] and graph mining [8]. Link mining is an instance

of multi-relational data mining (in its broadest sense); however, we use the term *link mining* to put an additional emphasis on the links—moving them up to first-class citizens in the data analysis endeavour.

Link mining encompasses a range of tasks including descriptive and predictive modelling. Both classification and clustering in linked relational domains require new data mining algorithms. But with the introduction of links, new tasks also come to light. Examples include predicting the numbers of links, predicting the type of link between two objects, inferring the existence of a link, inferring the identity of an object, finding co-references, and discovering sub graph patterns. We define these tasks and describe them in more detail in Section 3.

BACKGROUND

Probably the most famous example of exploiting link structure is the use of links to improve information retrieval results. Both the well-known page rank measure [35] and hubs and authority scores [27] are based on the link structure of the web. These algorithms are based on the citation relation between web pages. Recently, many algorithms have been proposed which examine other relations, for example, Dean and Hen zinger [9] proposed an algorithm based on co-citations to

find related web pages, or finer-grained representation of the web pages [5]. Richardson and Domingo's [40] combined content and link information with a relevance model to improve performance.A closely related line of work is hypertext and web page classification. This work has its roots in the information retrieval (IR) community. A hypertext collection has a rich structure that should be exploited to improve classification accuracy. In addition to words, hypertext has both incoming and outgoing links. Traditional IR document models do not make full use of the link structure of hypertext. In the web page classification problem, the web is viewed as a large directed graph. Our objective is to label the category of a web page, based on features of the current page and features of linked neighbours. With the use of linkage information, such as anchor text and neighbouring text around each incoming link, better categorization results can be achieved. Chakrabarti et al. [4] proposed a probabilistic model to utilize both text and linkage information to classify a database of patents and a small web collection. They showed that naively incorporating words from neighbouringpages reduces performance, while incorporating category information, such as hierarchical category prefixes, improves performance. Oh et al. [34] reported similar results on a collection of encyclopaedia articles: simply incorporating words from neighbouring documents was not helpful, while making use of the predicted class of neighbouring documents was helpful. These results indicate that simply assuming that link documents are on the same topic, and incorporating the features of linked neighbours, is not generally effective. Another approach to hypertext and link mining combines techniques from inductive logic programming with statistical learning algorithms to construct features from related documents. A pioneering example is the work of Slattery and Craven [43]. They proposed a model which goes beyond using words in a hypertext document making use of anchor text, neighbouring text, capitalized words and alphanumeric words. Using these statistical features and a relational rule learner based on FOIL [39], they proposed a combined model for text classification. Popescul et al. [38] also combined a relational learner with a logistic regression model to improve accuracy for document mining.

Other approaches to link mining identify certain types of hypertext regularities such as encyclopaedic regularity (in which linked objects typically have the same class) and co-citation regularity (in which linked objects do not share the same class, but objects that are cited by the same object tend to have the same class). Yang et al. [48] gave an in-depth investigation of the validity of these regularities across several datasets and using a range of classifiers. They found that the usefulness of the regularities varied, depending on both the dataset and the classifier being used.

Another link mining task that has received increasing attention is the identification of communities or groups, based on link structure. Gibson et al. [20] gave a survey of work in discovering Web communities. Kubica et al. [29] proposed a probabilistic model for link detection and modelling groups that makes use of demographic information and linkage information to infer group membership. Social and collaborative filtering has also been a focus of research that can be viewed as link mining. Kautz et al. [26] constructed social networks from Internet data and used the networks to guide users to experts who can answer their questions. Domingo's and Richarson [12] modelled the potential value of a customer, based on their network connections. Others have proposed generative probabilistic models for linked data. Cohn and Hofmann [7] proposed a probabilistic model for hypertext content and links. We also proposed a generative model for relational data, both content and links [17]. However, depending on the task, predictive models may be more appropriate. Examples of predictive modelling in relational domains include [44], [38], and [31].

LINK MINING TASKS

As mentioned in the introduction, link mining puts a new twist on some classic data mining tasks, and also poses new problems. Here we provide a (non-exhaustive) list of possible tasks. We illustrate each of them using the following domains as motivations:

Web page collection: In a web page collection, the objects are web pages, and links are in-links, out-links and cocitation links (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor text.

Bibliographic domain: In a bibliographic domain, the objects include papers, authors, institutions, journals and conferences. Links include the paper citations, authorship and co-authorship, affiliations, and the appears-in relation between a paper and a journal or conference.

Epidemiological Studies: In an epidemiology domain, the objects include patients, people they have come in contact with, and disease strains. Links represent contacts between people and which disease strain a person is infected with.

Link-Based Classification

The most straightforward upgrading of a classic data mining task to linked domains is *link-based classification*. In link based classification, we are interested in predicting the category of an object, based not just on its attributes, but on the links it participates in, and on attributes of objects linked by some path of edges.

An example of link-based classification that has received a fair amount of attention is web-page classification. In this problem, the goal is predict the category of a web page based on words on the page, links between pages, anchor text and other attributes of the pages and the links. In the bibliographic domain, an example of link-based classification is predicting the category of a paper, based on its citations, the papers that cite it, and co-citations (papers that are cited with this paper). In the epidemiology domain, an example is the task of predicting the disease type based on characteristics of the people (note the arbitrary possible prediction direction) or predicting the person's age, based on the disease they are infected with and the ages of the people they have been in contact with.

Link-based Cluster Analysis

The goal in cluster analysis is to find naturally occurring subclasses. This is done by segmenting the data into groups, where objects in a group are similar to each other and are very dissimilar from objects in different groups. Unlike classification, clustering is unsupervised and can be applied to discover *hidden patterns* from data. This makes it an ideal technique for applications such as scientific data exploration, information retrieval, computational biology, web log analysis, criminal analysis and many others.

There has been extensive research work on clustering in areas such as pattern recognition, statistics and machine learning. Hierarchical agglomerative clustering (HAC) and k-means are two of the most common clustering algorithms. Probabilistic model-based clustering is gaining increasing popularity [21 ; 45; 29]. All of these algorithms assume that each object is described by a fixed length attribute-value vector.

In the case of clustering linked data, even the definition of an element in a cluster is open to interpretation. We can cluster individual objects, collections of linked objects, or some other sub graph of the original. How do we compare the similarity of two of these elements or sub graphs, with potentially different structures? As this may necessitate tests for graph isomorphism, things will quickly become intractable. There has been surprisingly little work done on this type of link mining. Subdue [8] is the earliest line of research in this area. More recent approaches have been focused on efficiently finding frequently occurring patterns [30; 23]; these are largely inspired by the apriori algorithm [1] for mining frequently occurring patterns. One very interesting new approach is ANF [36], which attempts to compress a graph by approximating the neighbourhood function for each node. Examples of clustering in web page collections range from finding hubs (pages that point to lots of pages of the same category) to identifying mirror sites. Examples of clustering in the bibliographic domain include finding groups of authors that commonly publish together, and discovering research areas, based on common citations and common publication venues and discovering. An example of clustering in the epidemiology domain is finding patients with similar sets of contacts or diseases with similar transmission patterns. Next, we turn to some more specific tasks that arise in link mining. These can often be seen as special cases of link-based classification or link-based cluster analysis.

Identifying Link Type

There is a wide range of tasks related to predicting the existence of links. One of the simplest is predicting the type of link between two entities. For example, we may be trying to predict whether two people who know each other are family members, co-worker's, or acquaintances, or whether there is an adviser–advisee relationship between two co-authors.

The link type may be modelled in different ways. In some instances, the link type may simply be an attribute of the link. In this case, we may know the existence of a link between two entities, and we are simply interested in predicting its type. In our first example, perhaps we know there is some connection between two people, and we must predict whether it is a familial relation, a co-worker relation or acquaintance relation. In other instances, there may be different kinds of links. These may be different potential relationships between entities; in the second example, there are two possible relationships: a co-author relationship and an adviser–advisee relationship. We may want to make inferences about the existence of one kind of link, having observed another type of link.

A closely related task is predicting the *purpose* of a link. In a web page collection, the links between pages occur for different reasons. At the coarsest grain, links may be for navigational purposes or for advertising; it may be quite useful to distinguish between the two. The links may also indicate different relationships; the purpose of a link may be to refer to a professor's students, a student's friends, or a course's assignments.

Predicting Link Strength

Links may also have weights associated with them. In a web page collection, the weight may be interpreted as the authoritativeness of the incoming link, or its page rank. In an epidemiological domain, the strength of a link between people may be an indication of the length of their exposure.

Link Cardinality

There are many practical inferences that involve predicting the number of links between objects. The number of links is often a proxy for some more meaningful property whose semantics depend on the particular domain.

In a bibliographic domain, predicting the number of citations of a paper is an indication of the impact of a paper— papers with more citations are more likely to be seminal.
In a web collection, predicting the number of links to a page is an indication of its authoritativeness; predicting the number of links from a page is an indication that the page is a hub. The page rank measure is also clearly related to the number of links.

In an epidemiological setting, predicting the number of links between a patient and people with whom they have been in contact (their contacts) is an indication of the potential for disease transmission; predicting the number of links between a particular disease strain and people infected by it is an indication of the strain's virulence.

Note that link counts can be generalized to paths. A count of the number of paths between two objects may be significant.

Record Linkage

Another important concept in link mining is identity uncertainty [41; 37; 2]. In many practical problems, such as information extraction, duplication elimination and citation matching, objects may not have unique identifiers. The challenge is to determine when two similar-looking items in fact refer to the same object. This problem has been studied in statistics under the umbrella of record linkage [46; 47]; it has also been studied in the database community for the task of duplicate elimination [42].

In the link mining setting, it is important to take into account not just the similarity of objects based on their attributes, but also based on their links. In the bibliographic setting, this means taking into account the citations of a paper; note that as matches are identified, new matches may become apparent.

STATISTICAL MODELS FOR LINK MINING

Given the above collection of tasks, there are some unique challenges to applying staticallymodelling techniques. Here, we identify several; see also other papers in this volume, and papers in several recent workshops on learning statistical models from relational data [18; 19].

Logical vs. Statistical Dependences

The first challenge in link mining and multi-relational data mining is coherently handling two different types of dependence structures:

link structure - the logical relationships between objects

probabilistic dependency - the statistical relationship between attributes of objects. Typically we limit the probabilistic dependence to be among objects that are logically related. In learning statistical models for multi-relational data, we must not only search over probabilistic dependencies, as is standard in any type of statistical model selection problem, but potentially we must search over the different possible logical relationships between objects. This search over logical relationships has been a focus of research in inductive logic programming, and the methods and machinery developed in this community should be used to tackle this problem.

Feature Construction

A second challenge is feature construction in the multi-relational setting. The attributes of an object provide a basic description of the object. Traditional classification algorithms are based on these types of object features. In a link-based approach, it may also make sense to use attributes of linked objects. Further, if the links themselves have attributes, these may also be used. This is the idea behind propositionalization [15; 28]. However, as others have noted, simply flattening the relational neighbourhood around an object can be problematic. Several have noted that in hypertext domains, simply including words from neighbouring pages degrades classification performance [4; 34]. A further issue is how to deal appropriately with relationships that are not one-to-one. In this case, it may be appropriate to compute *aggregate* features over the set of related objects. We have found this works well for learning probabilistic relational models [16], but this approach may not always be appropriate.

Collective Classification

A third challenge is classification using a learned model. A learned link-based model specifies a distribution over link and content attributes, which may be correlated based on the links between them.

Intuitively, for linked objects, updating the category of one object can influence our inference about the categories of its linked neighbours. This requires a more complex classification algorithm than for a propositional learner. Iterative classification algorithms have been proposed for hypertext categorization [4; 34] and for relational learning [33; 45; 44]. The general approach of iterative classification has been studied in numerous fields, including relaxation-labelling in computer vision [22], inference in Markov random fields [6] and loopy belief propagation in Bayesian networks [32]. Some approaches make assumptions about the influence of the neighbour's categories (such as that linked objects have similar categories); we believe it is important to *learn* how the link distribution affects the category. As an example, this allows us to learn the notion of hubs – e.g., a computer science department homepage is likely to point to a lot of professor homepages.

Effective Use of Unlabeled Data

Recently there has been increased interest in learning using a mix of labelled and unlabelled data. General approaches include semi-supervised learning, co-training and transductive inference. There are some the unique ways in which unlabelleddata can be used to improve classification performance in relational domains. Just as in the case of the classical machine learning framework, in which there are no links among the data, unlabelled data can help us learn the distribution over object descriptions.

Links among the unlabelled data (or test set) can provide information that can help with classification.

Links between the labelled training data and unlabelled (test) data induce dependencies that should not be ignored.

Link Prediction

A fifth challenge is link discovery, or predicting the existence of links between objects. A range of the tasks that we have described fall under the category of link prediction. A difficulty here is that the prior probability of a link among any set of individuals is typically quite low. While we have had some success with simple probabilistic models of link existence [17] , we believe this is an area where there is much research to be done.

A further challenge is the discovery of common relational patterns or sub graphs; some progress has been made in this area [8; 30; 10]; however, this is an inherently difficult problem.

Object Identity

A final challenge is identity detection. How do we infer aliases, i.e., determine that two objects refer to the same individual? As mentioned earlier, some work has been done in this area by several research communities, but there is a great deal of room for additional work.

Another aspect of this challenge is whether our statistical models refer explicitly to individuals, or only to classes or categories of objects. In many cases, we'd like to model that a connection to a particular object or individual is highly predictive; on the other hand, if we'd like to have our models generalize and be applicable to new, unseen objects, we also have to be able to model with and reason about generic collections of objects.

CONCLUSION

There has been a growing interest in learning from linked data, which are described by a graph in which the nodes in the graph are objects and the edges/hyper-edges in the graph are links— or relations—between objects. Tasks include hypertext classification, segmentation, information extraction, searching and information retrieval, discovery of authorities and link discovery. Domains include the world-wide web, bibliographic citations, criminology and bio-informatics, to name just a few. Learning tasks range from predictive tasks, such as classification, to descriptive tasks, such as the discovery of frequently occurring sub-patterns. We have given a brief summary of some of the work in this area, and some of the challenges in link mining. Link mining is a promising new area where relational learning meets statistical modelling; we believe many new and interesting machine learning research problems lie at the intersection, and it is a research area "whose time has come" [11].

Acknowledgments

This paper was taken by us by referring many published papers under the acknowledgement of rules and regulations.

REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12 – 15 1994.

[2] M. Bilenko and R. J. Mooney. On evaluation and training-set construction for duplicate detection. under review.

[3] S. Chakrabarti. *Mining the Web*. Morgan Kaufman, 2002.

[4] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proc of SIGMOD-98*, 1998.

[5] S. Chakrabarti, M. Joshi, and V. Tawde. Enhanced topic distillation using text, markup tags, and hyperlinks. In *Research and Development in Information Retrieval*, pages 208–216, 2001.

[6] R. Chellappa and A. Jain. *Markov random fields: theory and applications*. Academic Press, Boston, 1993.

[7] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems 13*, 2001.

[8] D. Cook and L. Holder. Graph-based data mining. *IEEE Intelligent Systems*, 15(2):32–41, 2000.

[9] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, 31(11– 16):1467–1479, 1999.

[10] L. Dehaspe, H. Toivonen, and R. D. King. Finding frequent substructures in chemical compounds. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, editors, *4th International Conference on Knowledge Discovery and Data Mining*, pages 30–36. AAAI Press., 1998.

[11] P. Domingos. Prospects and challenges for multirelational data mining. *SIGKDD Explorations*, 2003. In this volume.

[12] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, 2001.

[13] S. Dzeroski and N. Lavrac, editors. *Relational Data Mining*. Kluwer, Berlin, 2001.

[14] R. Feldman. Link analysis: Current state of the art. In *KDD-02 Tutorial*, 2002.

[15] P. A. Flach and N. Lavrac. The role of feature construction in inductive rule learning. In *Proc. of the ICML2000 workshop on Attribute-Value and Relational Learning: crossing the boundaries*, 2000.

[16] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In S. Dzeroski and N. Lavrac, editors, *Relational Data Mining*, pages 307– 335. Kluwer, 2001.

[17] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models with link uncertainty. *Journal of Machine Learning Research*, 2002.

[18] L. Getoor and D. Jensen. *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*. AAAI Press, 2000.

[19] L. Getoor and D. Jensen. *Proc. IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*. AAAI Press, 2003.

[20] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[21] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

[22] R. Hummel and S. Zucker. On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):267–287 , 1983.

[23] A. Inokuchi, T. Washio, and H. Motoda. An aprioribased algorithm for mining frequent substructures from graph data. In *Principles of Data Mining and Knowledge Discovery*, pages 13–23, 2000.

[24] D. Jensen. Statistical challenges to inductive inference in linked data. In *Seventh International Workshop on Artificial Intelligence and Statistics*, 1999.

[25] D. Jensen and H. Goldberg. *AAAI Fall Symposium on AI and Link Analysis*. AAAI Press, 1998.

[26] H. Kautz, B. Selman, and M. Shah. Referral Web: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

COMPARATIVE STUDY ON PRIVACY PRESERVING THROUGH ANONYMIZATION BASED ON CLUSTERING ALGORITHMS OVER SOCIAL NETWORKS

[1]Asha Yesbhi.C [2]Murugesakumar.B
[1]M.Phil,(Research Scholar) [2]Assistant Professor
Department of Computer Science
Dr.SNS Rajalakshmi College of Arts and Science

ABSTRACT

Knowledge discovery on social network data helps in discovering current social trends and thus provide valuable findings that do good to the welfare of the general public. The researchers find that social networks play a vital role in examining people's behaviors. Before undergoing a research on social network data, the data needs to be anonymized to prevent potential re-identification attacks. Thus in this paper we are going to undergo a comparative study on clustering algorithms that are used to anonymize data over social networks.

Index Terms— Knowledge Discovery, Social Network, Clustering Algorithm, Anonymization

I. INTRODUCTION

To prevent privacy attacks on data over social network the first and foremost thing to be done is that data should be properly anonymized by taking into account the privacy model of the data and the utility of the data. Anonymization of relational data can be carried out using generalization and perturbation methods. Whereas privacy preservation on social network data is a new problem. Researches had taken place to develop privacy preservation methods among various strategies developed clustering based approach seems to be state-of-the-art anonymization methods on social network data. A clustering-based method clusters vertices and edges into groups and anonymizes a sub-graph into a super-vertex. In such a way, the details about individuals can be hidden properly.

II.CHALLENGES IN ANONYMIZING SOCIAL NETWORK DATA

Anonymizing relational data faces less challenge than anonymizing social network data because of the following functionality.

1. It is challenging to model background knowledge of adversaries and attacks about social network data. In a social network, many pieces of information can be used to identify component, such as labels of edges and vertices, neighborhood graphs, induce sub graphs, and their combinations. It is much more complicated and much more difficult than the relational case.

2. It is challenging to measure the information loss in anonymizing social network data . A social network consists of a set of edges and a set of vertices. It is hard to compare two social networks by comparing the edges and vertices individually. Two social network may differ in their network wise property such as betweeness ,connectivity ,diameter etc..Even though they have similar number of edges and vertices Thus, information loss and anonymization quality can be assessed by different methods .

3. At Last, it is challenging to devise anonymization methods for social network data than for relational data. Anonymizing a social network is difficult since changing labels of edges and vertices may affect the neighborhoods of other edges, and removing or adding edges and vertices may affect other edges and vertices as well as the properties of the network

III.CLUSTER- BASED ANONYMIZATION ALGORITHMS

In this section we would be discussing three main cluster based anonymization algorithm .They are

1. Bounded t-Means Algorithm
2. Union Split Algorithm
3. Sequence clustering Algorithm

3.1. Bounded t-Means Algorithm

It has been studied that classic clustering algorithm recurrently generate small or empty clusters specially when clustering high-dimensional datasets. To facilitate the realization of K-anonymity it is required that each cluster must possess the size at least K. This limitation to hold minimum size clustering led to the development of bounded t-means clustering method The bounded t-means clustering method is explained below.

1. Let t = [n/k]. Randomly choose t vertices to be cluster centers and represent them as vc1…vct. Denote the t clusters by c1,..ct. To begin with, all clusters are empty.

2. For each vertex v ∈ V:

(a) Cluster assignment Add vertex v to the nearest cluster ci according to a chosen distance Metric. We define the distance between two clusters to be

D (c1, c2) = D (vc1, vc2).

(b) Bumping If |ci| = k + 1, i.e. ci was full and already had k members, then perform the following procedure:

For each vertex u ∈ ci, compute the marginal cost g(u, ci, cj), where cj is the surrogate cluster of u. Bump the vertex u* with the lowest marginal cost to its surrogate cluster ci .

(c) Extra vertices If |V | is not a multiple of k, remainder vertices may be safely placed in their nearest cluster, without fear of violating the minimum-size constraint.

3. Cluster update for each cluster ci, a new cluster center v*ci is computed. If the new cluster centers are the same as for the previous iteration, the equilibrium has been reached and the algorithm terminates. Otherwise, repeat from Step 2.

3.2 .Union Split Algorithm

Both the t-means algorithm and bounded version does not guarantee to generate a globally optimal clustering solution. The property of t-means algorithm of choosing set of cluster centers randomly affects the clustering outcome. To avoid this inconsistency in clustering results, we design a new deterministic clustering algorithm, the union-split algorithm, which is described below.

1. Initialize each vertex to be in its own cluster.
2. Compute all pair-wise distances between cluster centers. For each cluster, maintain the next nearest cluster to it using a min-heap data structure.
3. While there exists an undersized cluster (< k members):

(a) Choose an undersized cluster c whose distance to its nearest cluster (full or undersized) is the shortest. Union cluster c with its nearest cluster.

(b) If the combined cluster c′ is overfull (size ≥ 2k),split it into two clusters each of size ≥ k. Splitting may be accomplished by finding the two vertices

(c) Update all relevant cluster distances.

4.When all clusters are full, stop.

3.3 Sequence Clustering Algorithm

The sequential clustering algorithm is a very efficient algorithm in terms of runtime as well as in terms of the utility of the output anonymization. We proceed to describe an adaptation of it for anonymizing social networks.

Input: A social network SN, an integer k.
Output: A clustering of SN into clusters of size ≥ k.
1.Choose a random partition C = {C1, . . . ,CT } of V intoT := ⌊N/k0⌋ clusters of sizes either k0 or k0 + 1.
2.For n = 1, . . . ,N do:
a) Let Ct be the cluster to which vn currently belongs.
b) For each of the other clusters, Cs, s≠ t, compute the difference in the information loss, Δn:t→s, if vn would move from Ct to Cs.
c) Let Cs0 be the cluster for which Δn:t→s is minimal.
d) If Ct is a singleton, move vn from Ct to Cs0 and remove cluster Ct.
e) Else, if Δn:t→s0 < 0, move vn from Ct to Cs0

3. If there exist clusters of size greater than k1, split each of them randomly into two equally-sized clusters.

4. If at least one node was moved during the last loop, go to Step 2.

5. While there exist clusters of size smaller than k, select one of them and unify it with the cluster which is closest.

6. Output the resulting clustering.

IV.COMPARATIVE STUDY OF CLUSTER BASED ANONYMIZATION ALGORITHMS

Based on the study made it is obvious that sequence clustering algorithm was an efficient algorithm both in terms of runtime as well as in terms of the utility of the output anonymization since they evaluate the information loss at each stage of clustering. The sequential clustering algorithm outperforms SaNGreeA algorithm as well they group dissimilar nodes on large social network graph. Whereas other two algorithms is an effective and general approaches to grouping similar nodes on large social network graphs for anonymization.

Algorithm	Advantages	Disadvantage
Bounded t-means clustering algorithm	i) With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small). ii) K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.	i)Difficulty in comparing quality of the clusters produced ii) Fixed number of clusters can make it difficult to predict what K should be. iii) Does not work well with non-globular clusters. iv) Different initial partitions can result in different final clusters.
Union Split Algorithm	i)It converges in a finite set of iteration ii) The complexity of the union-split clustering algorithm is O(n2 log n), where n represents the number of vertices in the original graph.	i)It does not work on grouping dissimilar nodes on large social network graph.
Sequence Clustering Algorithm	i)It is an effective algorithm in terms of runtime as well as in terms of the utility of the output anonymization ii)It evaluates at each stage during its operation the actual measure of information loss, since at each stage it has a full clustering of all nodes. iii) algorithms produce anonymizations by means of clustering with better utility than those achieved by existing algorithms.	i) The evaluation of cost function is complex as it needs to update the inter cluster costs for each candidate alternative clustering .and for each it considers $O(N/k)$ alternative cluster allocations

Table2.1:-The advantages and the disadvantages of three clusters based anonymization algorithm.

V.CONCLUSION

Thus from the observation on clustering based anonymization algorithms it is evident that social network graph does not consist of same number of vertices, edges or sub graphs all the time and they differ based on the properties of the network in order to obtain an effective runtime to achieve secured anonymization of social network data. For which the best suggested algorithm among the above three would be sequence clustering algorithm.

VI.REFERENCES

1.L. Backstrom, C. Dwork, and J. M. Kleinberg. Wherefore art though anonymized social networks, hidden patterns, and structural steganography. In *WWW*, pages 181–190, 2007.

2.F. Bonchi, A. Gionis, and T. Tassa. Identity obfuscation in graphs through the information theoretic lens. In *ICDE*, pages 924–935, 2011.

3. A. Campan and T. M. Truta. Data and structural k-anonymity in social networks. In *PinKDD*, pages 33–54, 2008.

4.J. Goldberger and T. Tassa. Efficient anonymizations with enhanced utility. *TDP*, 3:149–175, 2010.

5. S. Hanhij¨arvi, G. Garriga, and K. Puolamaki. Randomization techniques for graphs. In *SDM*, pages 780–791, 2009.

6. M. Hay, G. Miklau, D. Jensen, D. F. Towsley, and P. Weis. Resisting structural re-identification in anonymized social networks. In *PVLDB*,pages 102–114, 2008.

7. M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. *Uni. of Massachusetts Technical Report*, 07(19), 2007.

8. V. Iyengar. Transforming data to satisfy privacy constraints. In *ACMSIGKDD*,pages 279–288, 2002.

9. W. Jiang and C. Clifton. A secure distributed framework for achieving *k*-anonymity. *The VLDB Journal*, 15:316–333, 2006.

10.K. Liu, K. Das, T. Grandison, and H. Kargupta.Privacy-preserving data analysis on graphs and social networks. In H. Kargupta, J. Han, P. Yu, R. Motwani,and V. Kumar, editors, *Next Generation Data Mining*.CRC Press, 2008.

11.K. Liu and E. Terzi. Towards identity anonymization on graphs. In *Proceedings of the 2008 ACM SIGMOD In-ternational Conference on Management of Data (SIG-MOD'08)*, pages 93{106, New York, NY, USA, 2008.ACM Press.

12. L. Liu, J. Wang, J. Liu, and J. Zhang. Privacy preserving in social networks against sensitive edge disclosure.Technical Report Technical Report

CMIDA-HiPSCCS006-08, Department of Computer Science, University of Kentucky, KY, 2008.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Using the Computer Brain Cognitive Computing

Bharathi Anbarasan MCA, M. Phil
Department of Computer Applications
Dr. N.G.P. Arts and Science College
Coimbatore

ABSTRACT

Computers for quite some time now have been designed to think and work like a human being. Artificial intelligence uses the fundamental capability of computers to perform repeated mechanical computations, using fixed programming rules. However, the limitations of Artificial Intelligence led to the evolvement of Cognitive computing. Cognitive computing systems learn and interact naturally with people to extend what either humans or machine could do on their own. These systems will learn and interact to provide expert assistance to scientists, engineers, lawyers, and other professionals in a fraction of the time it now takes.

Index Terms—Artificial Intelligence, Cognitive Computing, Watson, Big Data

I. INTRODUCTION

Traditionally, computing has always been defined as a process of manipulation of data by a machine (computer) according to a set of instructions called a computer program. The program has an executable form that the computer can use directly to execute the instructions. The same program in its human-readable source code form, enables a programmer to study and develop the algorithm. Because the instructions can be carried out in different types of computers, a single set of source instructions converts to machine instructions according to the central processing unit type. Thus, a computer was defined to function based on instructions and not do anything more than that.

Later advances in technology over generations have led us to believe that computers can be designed to think and work like humans. Originally referred to as artificial intelligence, a field was founded on the claim that a central property of humans, intelligence - the sapience of *Homo sapiens* - can be so precisely described that it can be simulated by a machine.

From this initial introduction of the concept of Artificial intelligence, research has evolved in the field to develop systems which are able to think and interact with humans and consequently benefit the quality of human output rather than just remaining as an artificial system.

II. ARTIFICIAL INTELLIGENCE

The branch of computer science concerned with making computers behave like humans. The term was coined in 1956 by John McCarthy at the Massachusetts Institute of Technology.

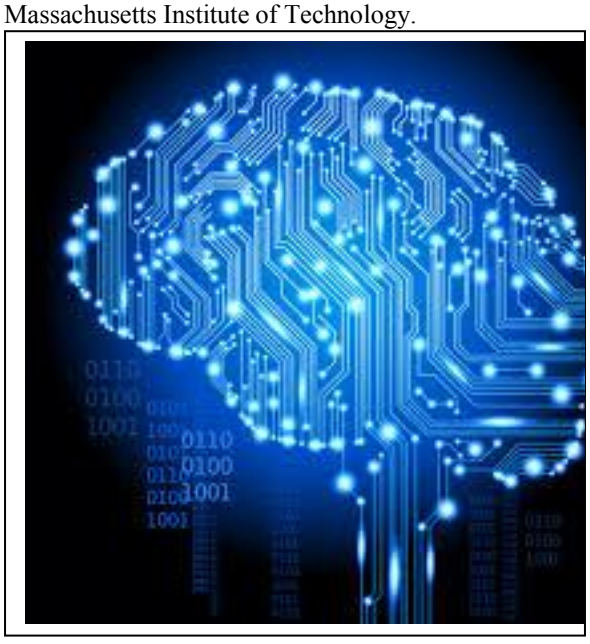


Fig. 1. Artificial Intelligence – The Human Machine

A. Why Artificial Intelligence?

Computers are fundamentally well suited to performing mechanical computations, using fixed programmed rules. This allows artificial machines to perform simple monotonous tasks efficiently and reliably, which humans are ill-suited to. For more complex problems, things get more difficult... Unlike humans, computers have trouble understanding specific situations, and adapting to new situations.

Artificial Intelligence aims to improve machine behaviour in tackling such complex tasks.

Together with this, much of AI research is allowing us to understand our intelligent behaviour. Humans have an interesting approach to problem-solving, based on abstract thought, high-level deliberative reasoning and pattern recognition. Artificial Intelligence can help us understand this process by recreating it, then potentially enabling us to enhance it beyond our current capabilities.

Artificial intelligence is applied in several fields including the following:

- **Games playing:** programming computers to play games such as chess and checkers
- **Expert systems :** programming computers to make decisions in real-life situations (for example, some expert systems help doctors diagnose diseases based on symptoms)
- **Natural language :** programming computers to understand natural human languages
- **Neural networks :** Systems that simulate intelligence by attempting to reproduce the types of physical connections that occur in animal brains
- **Robotics :** programming computers to *see* and *hear* and react to other sensory stimuli

B. Limitations of Artificial Intelligence

To date, all the traits of human intelligence have not been captured and applied together to spawn an intelligent artificial creature. Currently, Artificial Intelligence rather seems to focus on lucrative domain specific applications, which do not necessarily require the full extent of AI capabilities. This limit of machine intelligence is known to researchers as narrow intelligence.

Currently, no computers exhibit full artificial intelligence (that is, are able to simulate human behavior). The greatest advances have occurred in the field of games playing. The best computer chess programs are now capable of beating humans. In May, 1997, an IBM super-computer called *Deep Blue* defeated world chess champion Gary Kasparov in a chess match.

In the area of robotics, computers are now widely used in assembly plants, but they are capable only of very limited tasks. Robots have great difficulty identifying objects based on appearance or feel, and they still move and handle objects clumsily.

Natural-language processing offers the greatest potential rewards because it would allow people to interact with computers without needing any specialized knowledge. You could simply walk up to a computer and talk to it. Unfortunately, programming computers to understand natural languages has proved to be more difficult than originally thought. Some rudimentary translation systems that translate from one human language to another are in existence, but they are not nearly as good as human translators. There are also voice recognition systems that can convert spoken sounds into written words, but they do not *understand* what they are writing; they simply take dictation. Even these systems are quite limited -- you must speak slowly and distinctly.

In the early 1980s, expert systems were believed to represent the future of artificial intelligence and of computers in general. To date, however, they have not lived up to expectations. Many expert systems help human experts in such fields as medicine and engineering, but they are very expensive to produce and are helpful only in special situations.

Today, the hottest area of artificial intelligence is neural networks, which are proving successful in a number of disciplines such as voice recognition and natural-language processing.

III. COGNITIVE COMPUTING

Today, the term 'Cognitive computing' is used to refer to the development of computer systems modeled after the human brain. Researchers began to use the modern term (instead of Artificial Intelligence) in the 1990s, to indicate that the science was designed to teach computers to think like a human mind, rather than developing an artificial system. This type of computing integrates technology and biology in an attempt to re-engineer the brain, one of the most efficient and effective computers on Earth.

Cognitive computing systems learn and interact naturally with people to extend what either humans or machine could do on their own. They help human experts make better decisions by penetrating the complexity of Big Data.

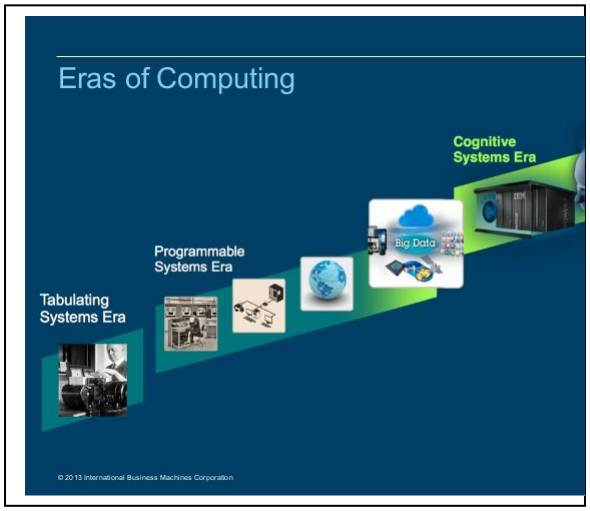


Fig. 2. Eras of Computing (Source: IBM)

IV. EXPERT SYSTEM VERSUS COGNITIVE COMPUTING

Cognitive computing systems get better over time as they build knowledge and learn a domain - its language and terminology, its processes and its preferred methods of interacting. Unlike expert systems of the past which required rules to be hard

coded into a system by a human expert, cognitive computers can process natural language and unstructured data and learn by experience, much in the same way humans do. While they'll have deep domain expertise, instead of replacing human experts, cognitive computers will act as a decision support system and help them make better decisions based on the best available data, whether in healthcare, finance or customer service. Watson's getting smarter with new capabilities to it. It's learning to reason, to think through things.

V. ARTIFICIAL INTELLIGENCE MEETS BUSINESS INTELLIGENCE

Big Data growth is accelerating as more of the world's activity is expressed digitally. Not only is it increasing in volume, but also in speed, variety and uncertainty. Most data now comes in unstructured forms such as video, images, symbols and natural language - a new computing model is needed in order for businesses to process and make sense of it, and enhance and extend the expertise of humans. Rather than being programmed to anticipate every possible answer or action needed to perform a function or set of tasks, cognitive computing systems are trained using **artificial intelligence (AI)** and **machine learning** algorithms to sense, predict, infer and, in some ways, think.

VI. HUMANS AND MACHINES WORKING TOGETHER

In traditional AI, humans are not part of the equation, yet in cognitive computing, humans and machines work together. To enable a natural interaction between them, cognitive computing systems use image and speech recognition as their eyes and ears to understand the world and interact more seamlessly with humans. It provides a feedback loop for machines and humans to learn from and teach one another. By using visual analytics and data visualization techniques, cognitive computers can display data in a visually compelling way that enlightens humans and helps them make decisions based on data.

Today, consumers continue to struggle to find and use actionable information. Traditional search systems deliver web pages, documents, video and audio to users when what they are really looking for is answers and advice. Cognitive Computing takes all of these information sources and distills them down to the important facts, events and relationships. Its natural language capabilities, hypothesis generation, cognitive analytics and machine-learning components then utilize these facts, events and relationships to answer questions in the same manner that a human would.

VII. CONTENT ANALYTICS AND DISCOVERY RESEARCH

Cognitive computing really comes down to three abilities

- The ability to perform deep natural language processing and analysis both for information ingestion and research, as well as to provide human-style communication.
- The ability to statistically generate and evaluate series of evidence-based hypotheses to be able to answer questions in a relevant and meaningful manner.
- The ability to adapt and learn from training, interaction with humans and outcomes related to hypotheses it generates.

These abilities find natural application in the field of Big Data Analytics where there is a need for being able to perform an enormous amount of processing over very large volumes of data in a very short period of time, to derive useful information. Big Data is increasing in volume, speed and uncertainty. It comes in unstructured forms such as video, image and text. To deal with this onslaught of information, new types of computing system are needed in order to understand, process and make sense of it all.

VIII. WATSON

The first cognitive computer was Watson, developed by IBM, which debuted in a television quiz show called 'Jeopardy!' The challenge for Watson was to answer questions posed in every nuance of natural language, such as puns, synonyms and homonyms, slang, and jargon. In the challenge, Watson bested the show's two greatest champions.

Watson was not connected to the Internet for the match. It only knew what it had amassed through years of persistent interaction and learning from a large set of unstructured knowledge. Using machine learning, statistical analysis and natural language processing to find and understand the clues in the questions, Watson then compared possible answers, by ranking its confidence in their accuracy, and responded – all in about three seconds.

Watson is ideal for applications where large amounts of information need to be ingested and understood, complex decisions are made and feedback is available to train the application to improve its decision making over time.

While Traditional computing systems are programmed to calculate rapidly and perform deterministic tasks, cognitive systems analyze information and draw insights from the analysis using probabilistic analytics. And they effectively continuously reprogram themselves based on what they learn from their interactions with data.



Fig. 3. The IBM Watson

IBM says cognitive computing systems like Watson are capable of understanding the subtleties, idiosyncrasies, idioms and nuance of human language by mimicking how humans reason and process information. Michael Rhodin, formerly senior vice president of IBM's Software Solutions Group, who has been tapped to lead the Watson Group, says it's a system that understands natural language," "It reads. When it reads a lot, it adapts and learns. It gets smarter. When you ask it questions, it will generate hypotheses--potential answers--with a degree of confidence."

A. Applications

Watson-powered apps slated for release recent year, including the following:

- MD Buyline, a provider of supply chain solutions for hospitals and healthcare systems, which is developing an app to allow clinical and financial users make real-time, informed decisions about medical device purchases.
- Welltok, a specialist in social health management, which is developing a mobile and web-based consumer app that will help users create "Intelligent Health Itineraries," which reward users for engaging in health behaviors.

IBM says that the power of this technology is really what it can do for everyone.

Fluid, a startup that builds online shopping experiences for major brands like Reebok and Brooks Brothers, which is building a personal shopper app, says

"The Watson Ecosystem is going to facilitate a wide range of these types of expert recommendation applications over the next several years and is going to continue to fuel research and innovation in using the tremendous growth in digital information to improve a wide range of processes from medical diagnosis to shopping to scientific research.

B. Watson Big Data Services

Out of the gate, IBM is offering three new services based on Watson's cognitive intelligence: Watson Discovery Advisor, Watson Analytics and Watson Explorer.

1) Watson Discovery Advisor

Watson Discovery Advisor is intended to help find the right questions in their data. IBM says it will revolutionize how industries like pharmaceutical and publishing conduct their research by reducing the time researchers need to formulate conclusions that can advance their work. For instance, with Discovery Advisor, Watson can pour through millions of articles, journals and studies, determine context, synthesize the data and help users pinpoint connections.

The Lancet used to explore how Watson's cognitive technologies can be used to help clinicians stay up-to-date on the medical knowledge necessary to give their patients the best possible care--an increasingly difficult task as medical data continues to grow at an exponential rate.

2) Watson Analytics

Watson Analytics allows users to explore big data insights through visual representations, without the need for advanced analytics training. IBM says the service removes common impediments in the data discovery process, giving business users the ability to quickly and independently uncover insights in their data. Watson Analytics prepares the data, surfaces the most important relationships and presents the results in an interactive visual format.

3) Watson Explorer

Watson Explorer is intended to help users across the enterprise uncover and share data-driven insights through a unified view that displays all of a user's data-driven information, as well as a framework for developing information-rich applications that deliver a comprehensive, contextually relevant view of topics.

DBS Bank announced last week that it will apply Watson, including Engagement Advisor, to its wealth management business to improve the advice and experience it delivers to affluent customers.

IX. THE FUTURE

Cognitive systems hold a great promise - a category of technologies that uses natural language processing and machine learning to enable people and machines to interact more naturally to extend and magnify human expertise and cognition. These systems will learn and interact to provide expert assistance to scientists, engineers, lawyers, and other professionals in a fraction of the time it now takes.

Far from replacing our thinking, cognitive systems will extend our cognition and free us to think more creatively. In so doing, they will speed innovations and ultimately help build a smarter planet.

REFERENCES

[1] www.33rdsquare.com,"The New Era of Cognitive Computing" , Web Publication, June 2013

[2] Rohan Pearce, "Watson and the future of Cognitive Computing" www.computerworld.com, July 2013

[3] IBM Research website, www.research.ibm.com

[4] Wikipedia articles on Artificial Intelligence, Cognitive Computing and Watson

[5] "What is Cognitive Computing?", wisegeek article, www.wisegeek.com

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A Study on Delay Tolerant Networks

[1]**Santhi.K** [2] **Deepa .J** Assistant Professor
Sri Ramakrishna College of Arts & Science Coimbatore

ABSTRACT

Delay-Tolerant Networks (DTNs) have the potential to interconnect devices in regions that current networking technology cannot reach. The idea is that an end-to-end connection may never be present. To make communication possible, intermediate nodes take custody of the data being transferred and forward it as the opportunity arises. Both links and nodes may be inherently unreliable and disconnections may be long-lived. To realize the DTN vision, routes must be found over multiple unreliable, intermittently-connected Many Researchers have investigated this fundamental challenge, particularly over the past five years. This paper reviewed the area of routing in delay-tolerant networks and presents a system for classifying the proposed routing strategies.

Index Terms - DTN, Flooding, Forwarding, Routing.

1. Introduction

Wired and wireless networks have enabled a wide range of devices to be interconnected over vast distances. For example, today it is possible to connect from a cell phone to millions of powerful servers around the world. The first and most important assumption is that an end-to-end connection exists from the source to the destination, possibly via multiple intermediaries. Delay-Tolerant Networking (DTN) is an attempt to extend the reach of networks. It promises to enable networks, sensor networks, mobile ad-hoc networks, and low-cost networks [1].

As an example of where these networks are useful, consider a classroom where each student has communication between "challenged" networks,

which includes deep space a laptop, but there is no network infrastructure [2]. One would like the students to collaborate on projects using the wireless network cards in the laptops, and also to communicate with the Internet. Delay-tolerant networking can make this happen as shown, in Figure 1. The laptops communicate with each other to exchange data. If the destination laptop is not present, the network stores the messages for it.

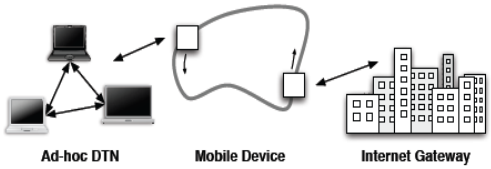


Figure 1: Laptops communicating with each other and the Internet via delay-tolerant networking

There are many other applications for delay-tolerant networks. In developing regions, applications range from education, health care and government services [3]. This allows very large files to be quickly and cheaply exchanged while small files and control messages are exchanged over a low bandwidth link. [4] This paper presents two properties that can be used to classify delay-tolerant routing strategies: replication and knowledge.

[5] Replication describes how a routing strategy relies on multiple copies of each message, and knowledge describes the information about the network which is used to make decisions. This review the routing strategies into two broad families, flooding and forwarding, based on their use of replication and knowledge.

2. Network Characteristics

DTN is composed of computing systems participating in the network, called nodes. One-way links connect some nodes together. These links may go up and down over time, due to mobility, failures, or other events. When the link is up, the source node has an opportunity to send data to the other end it is called a contact [1]. More than one contact may be available between a given pair of nodes. The contact schedule is the set of times when the contact will be available. The DTN architecture proposes to use this network by forwarding complete messages over each hop. These messages will be buffered at each intermediate node, potentially on non-volatile storage. This enables messages to wait until the next hop is available, which may be a long period of time. The amount of time for a message to be transferred from one node to another can be divided into four components: waiting time, queuing time, transmission delay, and propagation delay [6].

2.1. Challenges

Delay-tolerant networks present many challenges that are not present in traditional networks. Many stem from the need to deal with disconnections, which directly impacts routing and forwarding. There are secondary problems that routing strategies may need to be aware of, such as dealing with limited resources.

anywhere from seconds to days whereas the others are

typically much shorter. Thus, one of the most important characteristics of a DTN is the contact schedule, which depends strongly on the application area under consideration. Contact schedules can be placed on an approximate spectrum based on how predictable they are, as shown in Figure 2. These networks are widely studied in the ad-hoc networking community because the models are simple to work with. This spectrum is similar to the one presented in [7].

`

2.1.2 Contact Capacity

A question that is closely related to the contact schedule is how much data that can be exchanged between two nodes. This depends on both the link technology and the duration of the contact. If the volume of traffic is very small compared to the capacity of contacts in the network, then this is a reasonable approach. However, if the volume of traffic increases due to a large number of users, or due to large messages being exchanged, the contact capacity becomes very important [8].

3. Strategy Families

DTN routing strategies classified into two families based on which property a strategy uses in order to find the destination. The families are flooding strategies, which rely primarily on replicating messages to enough nodes so the destination receives it, and forwarding strategies, which rely on knowledge about the network to select the best path to the destination.

2.1.3 Buffer Space

In order to cope with long disconnections, messages must be buffered for long periods of time. This means that intermediate routers require enough buffer space to store all the messages that are waiting for future communication opportunities [9]. From one point of view, this means that intermediate routers require buffer space proportional to demand.

2.1.4 Processing Power

These devices may be very small, and similarly have small processing capability, in terms of CPU and memory. These nodes will not be capable of running complex routing protocols.

2.1.5 Energy

Some nodes in delay-tolerant networks may have limited energy supplies either because they are mobile, or they are in a location that cannot easily be connected to the power grid. Routing consumes energy by sending, receiving and storing messages, and by performing computation [11]. Hence, routing strategies that send fewer bytes and perform less computation will be more energy efficient.

3.1 Flooding Strategies

Strategies in the flooding family deliver multiple copies of each message to a set of nodes, called relays. The relays store the messages until they connect with the destination, at which point the message is delivered [12]. Message replication is then used to increase the probability that the message gets delivered.

3.1.1 Direct Contact

This strategy waits until the source comes into contact with the destination before forwarding the data. This is the degenerate case of the flooding family, where the set of relays contains only the destination. It can also be considered a degenerate case of the forwarding family, where it always selects the direct path between the source and the destination. Grossglauser and Tse showed that in their mobile ad-hoc network scenario, this strategy has a capacity that approaches zero as the number of nodes increases [13].

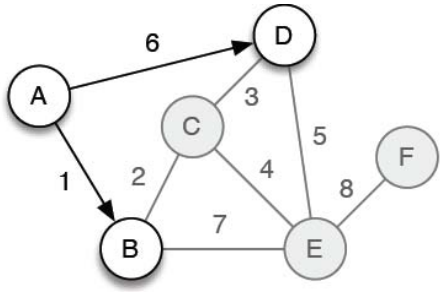


Figure 3: Direct Contact routing example

3.1.2 Two-Hop Relay

In this strategy, the source copies the message to the first n nodes that it contacts. The source and the relays hold the message and deliver it to the destination. Since there are now n+1 copies of the message in the network, more bandwidth and storage are consumed.

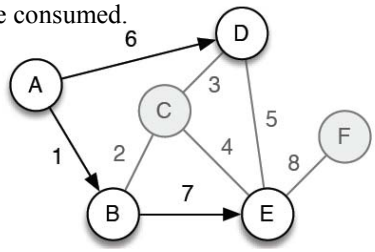
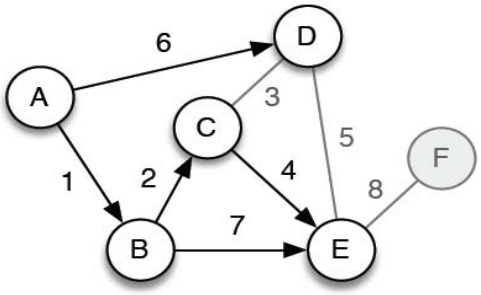


Figure 4: Two-hop routing example

3.1.3 Tree-Based Flooding

Tree-Based Flooding strategies extend two-hop relay by distributing the task of making copies to other nodes. When a message is copied to a relay, there an indication of how many copies the relay should make. [14] This is called Tree-Based Flooding because the set of relays forms a tree of nodes rooted at the source. Two-hop relay can be viewed as Tree-Based Flooding with a depth of one.

Figure 5: Tree-Based Flooding Example



3.2 Forwarding Strategies

They use network topology information to select the best path, and the message is then forwarded from node to node along this path. A path can be found using location-based routing, assigning metrics to nodes or by assigning metrics to links[16]. Some of these approaches have been explored in wired and multi-hop wireless networks.

4. Conclusion

Delay-tolerant networks are a promising new development in network research, that offer the hope of connecting people and devices that were unable to communicate, or could do so only at great cost [15]. In this paper to classify them according to two key properties: replication and knowledge.

First, to achieve a high delivery ratio with low resource consumption, hybrid techniques that rely on both knowledge about the topology and replication will be required.

Second, in cases where message volume is low, simple epidemic routing works extremely well [17].

5.REFERENCES

[1] S. Burleigh, A. Hooke, L. Torgerson, K. Fall, V. Cerf, B. Durst, K. Scott, and H. Weiss, "Delay-tolerant networking: an approach to interplanetary internet,"IEEE Communications Magazine, vol. 41, pp. 128– 136, June 2003.

[2] S. Jain, K. Fall, and R. Patra, "Routing in a delay tolerant network," in Proceedings of ACM SIGCOMM, vol. 34, pp. 145–158, ACM Press, October 2004.

[3] D. Marasigan and P. Rommel, "MV routing and capacity building in disruption tolerant networks," in Proceedings of IEEE INFOCOM, vol. 1, pp.

398– 408,March 2005.

[4] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "Maxprop: Routing for vehicle-based disruption-tolerant networking," in Proceedings of IEEE INFOCOM, April 2006.

[5] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft ,and C. Diot, "Pocket switched networks and human mobility in conference environments," in Proceedings of the ACM SIGCOMM Workshop on Delay-Tolerant Networking (WDTN'05), pp. 244–251, August 2005.

[6] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey," IEEE Wireless Communications, vol. 11, no. 6, pp. 6– 28, 2004.

[7] H. Jun, M. H. Ammar, and E. W. Zegura, "Power management in delay tolerant networks: A framework and knowledge-based mechanisms," in Proceedings of IEEE SECON, September 2005.

[8] R. H. Frenkiel, B. R. Badrinath, J. Borres, and R. D. Yates, "The infostations challenge: balancing cost and ubiquity in delivering wireless data," IEEE Personal Communications, vol. 7, no. 2, pp. 66–71, 2000.

[9] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of ad hoc wireless networks," IEEE/ACM Transactions on Networking, vol. 10, pp. 477–486, August 2002.

[10] A. Vahdat and D. Becker, "Epidemic routing for partially-connected ad hoc networks," Tech. Rep. CS-2000-06, Duke University, July 2000.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Blocking Records Through Tb Method For Large Data Warehouse

Dharani.R[1], Jayanth.D[2]

[1]Research Scholar, [2]Assistant Professor,

Department of Computer Science, Bharathidasan College of Arts and Science, Chithode, Erode,

ABSTRACT

Record linkage is a critical problem in duplicate data elimination. It is used to detect and eliminate duplicate data. The elimination of duplicate data will increase the quality of data. Record Linkage problem will take high computational cost because of the large number of record comparisons. The comparison of records is inefficient in large data warehouses. Blocking methods are used to group the records to minimize the number of record comparisons. This paper explains the existing blocking methods and its comparison and discusses the selection of token-based blocking key for record comparisons.

Keywords: Data Warehouse, Record Linkage, Token, Blocking Records, Record Comparisons, Duplicate Data

1. Introduction

Record linkage [RL] is the process of matching records across data sets that refer to the same entity. Theoretically, RL compares all records in the data sets under analysis in order to decide which record belongs to the same individual. In practice, since the size of data sets is usually very large, comparing all the records between them becomes unfeasible. Therefore, RL resorts to blocking methods that try to gather all the records that present a potential resemblance, only applying RL within each block. Typically, blocking methods are based on a common attribute without errors. Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Clustering method is also known as Blocking method in data cleaning for duplicate detection. Blocking typically refers to the procedure of subdividing data bases into a set of mutually exclusive subsets (blocks) under the assumption that no matches occur across different blocks. Although blocking can substantially increase the speed of the comparison process, it can also lead to an increased number of false mismatches due to the misuse of blocking key and the size of the window. There are two main goals of blocking. First, the number of candidate matches generated should be small to minimize the number of detailed comparisons in the record linkage step. Second, the candidate set should not leave out any possible true matches, since only record pairs in the candidate set are examined in detail during record linkage. These blocking goals represent a trade off. On the one hand, the goal of record linkage is to find all matching records, but the process also needs to scale. This makes blocking a challenging problem. The main purpose of this paper is to substantially reduce the probability of false mismatches and to increase the running time, by reducing the number of comparisons. This paper addresses existing blocking methods in Record Linkage providing a summary and comparison of the same. Also, this paper introduces the efficient token-based blocking key selection blocking method. The blocking key is important in this blocking method to gather resemblance records.

2 Clustering / Blocking Methods2.1 Standard Blocking

The Standard Blocking (SB) method cluster records into from the record attributes in ch ata set. Composed of more than one attribute. The resulting total number of record pair comparisons of the standard blocking is O(n2/b).

2.2 Sorted Neighborhood Method (SNM)

One of the most common duplicate detection methods is the Sorted Neighborhood method. The Sorted Neighborhood method sorts the records based on a sorting key [SK] and then moves a window called Sliding window (SW) of fixed size w sequentially over the sorted records. Records within the window are then paired with each other and included in the candidate record pair list. The use of the window limits the number of possible record pair comparisons for each record to

2w−1. The resulting total number of record pair comparisons of the sorted neighborhood method is O(wn).

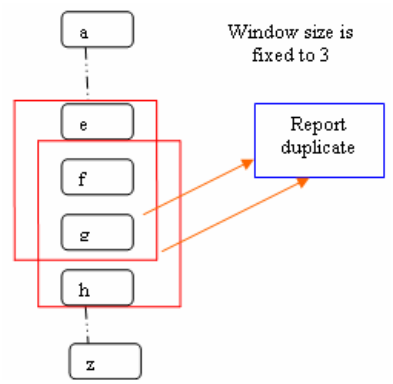


Fig 1: Sorted Neighborhood Method (SNM)

One problem with the sorted neighborhood method is, if a number of records larger than the window size have the same value in a sorting key. Similar to standard blocking, it is advantageous to do several passes (iterations) with different sorting keys and a smaller window size than one pass only with a large window size. The effectiveness of this approach depends on the quality of key chosen to do the sorting. The multi-pass sorted neighborhood method involves using multiple passes of the initial sorted neighborhood method. It is executed in several independent runs of SNM, each time using a different key and a relatively small window. In general, no single key will be sufficient to catch all matching records. During each pair, the three steps - create keys, sort data and merge – are performed selecting a different key field for each pass. Once the duplicate records are identified during a pass one of the duplicates is eliminated from the data source. This method provides better results than the SNM method.

2.4 Clustering SNM

Clustering SNM, first partitions the database into independent clusters using a key extracted from the data [3]. Then SNM is applied to each individual cluster independently. There are two phases in Clustering SNM.

1. Cluster data

Scan the records in sequence to extract a key for each record; then using the extracted key, partition the records into independent subsets of data (clusters).

2. SNM

Apply SNM independently on each cluster. The key does not need to be recomputed, the key extracted above can be used for sorting.

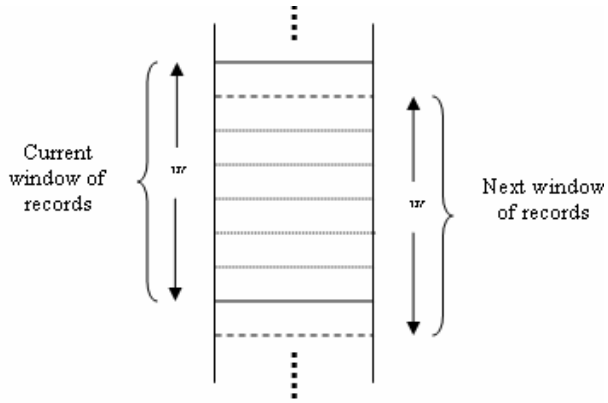


Fig. 2: Sliding Window

2.5 Duplicate Elimination SNM (DE-SNM)

Duplicate Elimination SNM sorts the records on a chosen key and then divides the sorted records into two lists. 1) a duplicate list 2) a non-duplicate list. A small window scan is first performed on the duplicate list to find the list of matched and unmatched records. The list of unmatched records is merged with the original non-duplicate list and a second window scan is performed. Using Sorted Neighborhood method the exact or very closely matching duplicate records with equivalent keys can be identified and it would likely be more efficient to find these records first during the sorting phase. This method removes other duplicate records and retains only one representative member of this set of duplicates by the rule set against other records with different keys. DE- SNM does not contribute much on the improvement of accuracy of SNM. The benefit of DE-SNM is that it runs faster than SNM under the same window size, especially for the databases that are heavily dirty. If the number of records in duplicate list is large, DE-SNM will run faster than SNM.

2.6 SNM – Priority Queue

The algorithm scans the database sequentially and determines whether each record is scanned or not and also determines whether a member of the cluster is represented in priority queue. This queue contains a fixed number of record sets, which contain similar records that can be paired as candidate duplicates. The sorted list of all records is sequentially scanned and every record

Rj is compared with the members Ri in the record set with the highest priority in the queue. If the comparison yields a distance below some threshold T1, the record Rj is included into the set of Ri, if the distance is higher than some threshold T2 this record set is skipped and the comparison is continued with the next highest priority set in the queue. If no set is found at all, a new singleton set is put into the queue. In order to reduce the number of comparisons, the record sets are pruned, i.e. they only contain ”representative” members of the set, in particular, non-representative members are those, which are very similar to other members. However, the authors do not make it clear how to find the two thresholds and how to decide if a record is “representative”.

2.7 K-way Sorting Method

The Sorted Neighborhood Method is not effective where the data source does not contain a primary key field of reference. The k-way sort duplicate detection method is designed to effectively detect duplicates in data sources that do not have any uniquely distinguishing data fields. This K-way sorting method follows the following steps.

The steps are

1. Let k be the number of columns to be used for sorting

2. Select the k-most meaningful combinations of sort keys based on the k selected columns.

3. Assign a record identifier to each record.

4. Sort records based on the selected sort key combination

5. For each sorted set of data, compare adjacent rows within a given window size

6. Draw k graphs for each sort

7. Examine the k graphs collectively. If matches occur between some records or identifier exceeds certain threshold, then it should be mapped into the summation graph.

8. The summation graph should handle transitive closure.

2.8 Bigram Indexing

The Bigram Indexing (BI) method as implemented in the Febrl record linkage system allows for fuzzy blocking. The basic idea is that the blocking key values are converted into a list of bigrams (substrings containing two characters) and sub-lists of all possible permutations will be built using a threshold (between 0.0 and 1.0). The resulting bigram lists are sorted and inserted into an inverted index, which will be used to retrieve the corresponding record numbers in a block. The number of sub-lists created for a blocking key value both depends on the length of the value and the threshold. The result of bigram list is smaller blocks will be produced in the inverted index if the value of threshold is low and the sub-list is small with per blocking key value. In the information retrieval field, bigram indexing has been found to be robust to small typographical errors in documents. Like standard blocking, the number of record pair comparisons with two data sets with n records each and b blocks, all containing the same number of records is O(n2 b). The number of blocks b will be much larger in bigram indexing.

2.9 Canopy Clustering with TF-IDF

Canopy Clustering with TF-IDF (Term Frequency- Inverse Document Frequency) forms blocks of records based on those records placed in the same canopy cluster. A canopy cluster is formed by choosing a record at random from candidate set of records. This canopy clustering method uses the TF-IDF distance metric instead of bigrams [1]. The total number of record pair comparisons resulting from canopy clustering is O(fn2/c) where n is the total number of records, f is the average number of canopy and c is the number of canopies. The threshold parameter should be set so that f is small and c is large, in order to reduce the amount of computation. However, if f is too small, then the method will not be able to detect typographical errors.

2.10 SNM-IN

The main difference between SNM and SNM-IN is as follows: while SNM compares the new record entering the current window with all previous records in the window, SNM-IN will first check whether the new record with previous records are duplicate or non-duplicate with the new record, there is no need to compare them with the new record. Thus, with Lower bound and Upper bound, SNM-IN will reduce a lot of comparisons.

2.11 SNM-INOUT

SNM-IN follows SNM with only one anchor record and SNM-INOUT follows SNM with two anchor records. One anchor record is an in Anchor record and the other is out Anchor record. SNM-INOUT has some changes in Merge

phase of SNM method. When a new record enters the current window, it is first compared with the two anchor records, which introduces one more comparison than SNM-IN. However, in SNM-INOUT, for each record in the current window, the chance of determining the new record as duplicate record or not is increased, with produce more chance to reduce comparisons than SNM-IN does. Since the out Anchor is outside the window and the last record in the window will compare with it, SNM-INOUT will obtain a few more duplicate pairs than SNM-IN if both run at the same window size.

2.12 Disjunctive Blocking

Disjunctive blocking selects record pairs that are placed in the same block by at least one blocking predicate from the selected subset of predicates. This strategy can be viewed as selecting pairs for which a disjunction of predicates evaluates to true. The blocking function is trained by selecting a subset of blocking predicates.

2.13 Disjunctive Normal Form (DNF) Blocking

DNF blocking selects object pairs that satisfy at least one conjunction of blocking predicates from a selected subset of conjunctions. This strategy can be viewed as selecting pairs for which a disjunction of predicate conjunctions evaluates to true. The blocking function is trained by constructing the DNF formula over the blocking predicates.

2.14 Fuzzy Blocking

In the standard blocking method, the blocking keys are selected from the noisy attributes. The fuzzy blocking method substitutes the blocking key by the results of an OWA (Ordered Weighting Averaging) operator. This fuzzy blocking method follows the following steps:

1. A data in the data set is normalized

2. A Fuzzy quantifier is selected

3. OWA operator with the fuzzy quantifier

4. The OWA result is rounded and the record is blocked using this value as a blocking key.

The fuzzy blocking method has two parameters:

The OWA quantifier that decides the way to aggregate, and the rounding method that decides the number of blocks to do.

2.15 Adaptive Sorted Neighborhood Method (A-SNM)

Adaptive Sorted Neighborhood Method uses the original SNM approach to slide a fix-sized window n - w + 1 times, to generate n-w+1 number of blocks when applied to n records. Note that these generated blocks are overlapping each other, and fix-sized. However, in the adaptive SNM, the “window" becomes only an interim tool to generate final blocks, and all generated blocks will have different (thus adaptive) sizes.

2.15.1 Incrementally Adaptive SNM (IA-SNM)

The basic idea is to measure whether records within a small neighborhood are close/sparse and if there are rooms to grow/shrink in the window, then the window size is increased/decreased dynamically. In order to measure the record distribution within a window, it seems that we need to measure the distances between all the records in the window. If the distance between the first and last record satisfies dist(r1,rw1), where _ is the distance threshold. This distance indicates that records within the current window are close to each other, so there is still room to enlarge the window size to find more potential duplicate records. Otherwise the window should be retrenched.

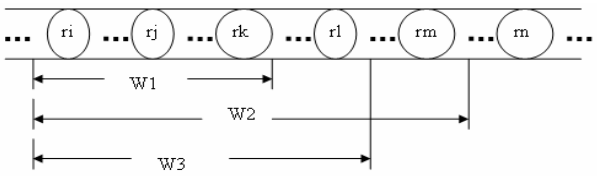


Fig 3: Incrementally-adaptive SNM (IA-SNM)

2.15.2 Accumulatively Adaptive SNM (AASNM)

Accumulatively Adaptive SNM is another scheme that can adaptively find various sized blocks. Unlike IA-SNM, which does not explicitly search for boundary pairs to find blocks, the goal of this method is to quickly and accurately find all boundary pairs. As in IA-SNM, AA-SNM measures the distance between the first and the last record in the current window W1. If the distance is less than or equal to a threshold, all the records in window Wm1 can be determined as potential duplicates to each other and should be grouped into the same block. It also suggests that the boundary pair is not in the current window and there are more potential duplicates outside windowWm1. Therefore, the window needs to enlarge the

neighborhood of comparison to include more potential duplicates and to search for the boundary pair. This can be achieved by either enlarging the window size, or by moving the window forward and connecting the results of consecutive windows later. The second option is used in AA-SNM. In the retrenchment phase, the binary comparison is done to reduce the time.

3 Comparison of Blocking Methods

In detection methods, the most reliable way is to compare every record with every other record. Obviously this method guarantees that all potential duplicate records are compared and then provides the best accuracy. However, the time complexity of this method is quadratic. It takes N(N-1)/2 comparisons if the database has N records, which will take very long time to execute when N is large. Thus it is only suitable for small databases and is definitely impracticable and infeasible for large databases. Therefore, for large databases, approximate detection algorithms that take far less comparisons (e.g., O(N) comparisons) are required. The window size used in SNM is an important parameter that affects the performance. If the window size is large then accuracy increases slowly but the time increases fast. The duplicate records missed in one pass would be defeated in another pass with different key, so multi-pass increases correctness. Multi-pass SNM can drastically improve the accuracy of the results of only one run of SNM with varying large windows. Particularly, only a small window size is needed for the multi-pass SNM to obtain high accuracy, while no single run with a key can produce comparable accuracy results with a large window. One issue in multi-pass SNM is that it employs transitive closure to increase the number of duplicate records. The transitive closure allows duplicate records to be detected even without being in the same window during an individual window scan. The multi-pass SNM also increases the number of false positives. Priority Queue can save some unnecessary comparisons taken by SNM by computing the transitive closure online. Priority Queue may be faster than SNM but cannot improve the accuracy under the same conditions with SNM. In addition, the performance of priority queue depends on the degree of dirtiness of databases. For clean and slightly dirty databases priority queue does not provide any help. But for dirty databases, priority queue is much faster. The more dirtier the database is, the more time it can save. Like priority queue, DE-SNM can also run faster than SNM for dirty databases, but DE-SNM will decrease the accuracy. Clustering SNM is an alternative method. As the name shows, clustering SNM does one even loser clustering before applying SNM. The clustering SNM is faster than SNM for very large databases but it may decrease the accuracy as well. Further, clustering SNM is suitable for parallel implementation. The bigram indexing needs the inverted index for the bigram sub-lists, where as canopy clustering needs a special index for finding all records that, are within the threshold of a distance measure, which depends on the used distance measure. The k-way sort duplicate detection method is designed to effectively detect duplicates in data sources that do not have any uniquely distinguishing data fields . Disjunctive and DNF blocking can then be performed by iterating over the lists for each block in the inverted indices and returning all pairs of records that co-occur in at least one such list. DNF blocking is more accurate than disjunctive blocking at 100% recall. SNM and SNM-IN obtain exactly the same duplicate result, and SNM-INOUT obtains slightly more duplicate pairs than SNM-IN does. In SNM-IN and SNM INOUT, the number of comparisons is saved with different window sizes. SNM-INOUT will take more comparisons than SNM-IN because SNM-INOUT uses two anchor records. IA-SNM and AA-SNM are robust to the variation in the distribution of duplicates. These both methods are adaptive to the variance of block sizes. AA-SNM has better performance than IA-SNM. SNM blocking method gives better performance than other methods like shrinking or expanding the window size based on the records. Finally, all the blocking methods are effective in blocking records but the quality of the data, the time taken for the record comparison and false positive rate are different in all the existing blocking methods. The false positive rate is reduced by the dynamic changes of window size and by the usage of an effective blocking key.

4 Token-Based Blocking key

Blocking is a mechanism used in record linkage to reduce the number of pair comparisons. A database (set of records) is divided into smaller blocks by blocking key values. Instead of comparing every possible pair that can be formed by records in the database, one will need to compare those pairs whose records belong to the same block. A blocking key is a pre-defined set of positions. A requirement for a good blocking key is that, the more the possibility that two records are duplicate, the more likely they are in the same block. The automatic selection of good blocking keys is very important to bring duplicate records together. In the existing method the first three or four characters are selected from the field value of single attribute which is having high power discrimination. This way of blocking key formation is not efficient to bring all the duplicate records

together. In some other existing method of blocking, the first character of selected attribute field value is selected as blocking key. Therefore, the selection of attribute is very important in the selection of blocking key. These high power discrimination attributes will be selected based on some criteria. This paper proposes attribute selection method for the data cleaning process. There are three criteria that are used to identify relevant attributes for further data cleaning process. The three criteria are: (a) Identifying key attributes, (b) classifying attributes with high distinct value and low missing value and (c) measurement types of the attributes.

a) Identifying key attributes

The key is an attribute or a set of attributes that uniquely identify a specific instance of the table. Every table in the data model must have a primary key whose values uniquely identify instances of the entity. The key may be primary key, candidate key, foreign key or composite key.

b) Classifying distinct and missing values

Missing character values are always same no matter whether it is expressed as one blank, or more than one blanks. Obviously, missing character values are not the smallest strings. Distinct is used to retrieve number of rows that have unique values for each attribute. The distinct value is used to calculate an Identification power of the attribute. So the distinct value is very much important in attribute selection. An Identification power of the attribute (ipj) is used to evaluate the discriminating power of record attributes.

$$\text{Identification power of the attribute j (ipj)} = \frac{\text{Number of distinct equivalence classes on the total of record}}{\text{Total number of records}}$$

c) Classifying types of attributes

There are four types of attributes: nominal, ordinal, interval and ratio. Different criteria are given for each attribute type. The value of these measurement types are also considered for the attribute selection. The data cleaning using numeric data will not be effective. Categorical data is efficient for the data cleaning process. The selected attributes will be used as the blocking key. If the window size is small, then the number of comparisons will be reduced but the false-positive value will be increased. If the window size is large then the number of comparisons will be increased but the false-positive value will be reduced. For this kind of problem, the calculation of distinct and missing value is very important in the selection of blocking key. The numeric blocking key will not be effective in blocking the record for comparison. So, the identification of measurement type and type of the attribute is important in the selection of blocking key. Token is formed for the selected attribute field value. Different token formation rules are followed for different kind of data. The data may be numeric, alphanumeric or alphabetic. The rules are given in the following steps.

a. Numeric Tokens: This numeric token formation rule is suitable for phone number, social security number, street number, apartment number, and so on. First, it removes the unimportant characters and converts the character to numeric. Finally, groups the number to keep together as one token.

b. Alphabetic Tokens: This alphabetic token formation rule is well suited for the names such as contact name, customer name, producer name, book title, and so on. First, it expands the abbreviations and removes the unimportant and stopping characters. Finally, takes the first character from each word, sorts the selected character then groups together as one token [16].

c. Alphanumeric Tokens: This alphanumeric token rule is suited for address, product code and so on. First, it splits alphanumeric into numeric and alphabetic, sorts the divided token and then groups numeric and alphabetic separately. Finally, the tokens in the field are grouped together to get token as one field.

This proposed work uses the existing blocking key approaches as well as it uses token based blocking key. The token is formed for each selected attribute field values before blocking the records. Then the blocking key is created from the token field by selecting the first character of each token filed. The token based blocking key is efficient to block the records. While blocking the records, the similarity function is used to group the records based on the threshold value. The proposed work will be efficient to reduce the false positive.

5. Conclusion

In this paper, existing blocking methods are presented for the blocking records to minimize the time and number of comparison in the record linkage. The selection of the most suitable blocking key (parameter) for the blocking method is addressed in this paper. Dynamically adjusting the blocking key for the blocking method will be effective in record linkage algorithms during the

execution time. The blocking key is selected based on the type of the data and usage of the data in the data warehouse. The dynamically adjusting blocking key and token based blocking key as well as the dynamic window size SNM method will be used in this proposed work. Future directions of this work include agents in tuning parameter or everything is set dynamically for the blocking method without human intervention to yield better performance. However, in most real world problems where expert knowledge is hard to obtain, it is helpful to have methods that can automatically choose reasonable parameters for us.

ACKNOWLEDGMENT

Dharani.R is thankful to Prof. D.Jayanth, Asst. Professor, Department of Computer Science, Bharathidasan College of Arts and Science,Erode, for his constant support and helping out with the preparation of this paper. She is also thankful to the Principal, Bharathidasan College of Arts and Science, Erode and Prof. M.Inbavel, HOD, Department of Computer Science, Bharathidasan College of Arts and Science,Erode for being a constant source of inspiration.

REFERENCE

[1] Rohan Baxter, Peter Christen and Tim Churches (2003) Record Linkage and Object Consolidation at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC.

[2] Chen Shengxin (2002) Morgan Kaufmann Publishers Language: ENGLISH. 16x24 Hardback, 21-26.

[3] Feekin A. and Chen Z. (2000) Proceeding ACM SAC Conference, 323-327.

[4] Li Zhao, Sung Sam Yuan, Sun Peng, Tok Wang Ling (2002) Proceedings of the 13th International Conference on Database and Expert Systems Applications, Publisher- Springer-Verlag, 2453, 484 – 493.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Categorizing Information Retrieval Query Using Classification Tree Algorithm

Dr. S. Vijayarani[1], Mr.S.Dhayanand[2], Ms. N. Nithya[3],

Assistant Professor[1], M.Phil Research Scholars[2, 3],
Department of Computer Science, School of Computer Science and Engineering,
Bharathiar University, Coimbatore, Tamilnadu, India[1,2,3].

ABSTRACT

Data mining is the process to extract information from the large database. Various data mining techniques are clustering, classification, association analysis, regression, summarization, time series analysis and sequence analysis, etc. Data mining techniques plays a vital role in different domains such as text mining, graph mining, multimedia mining, web mining, etc. Text mining is used for analyzing the information in the text data bases. Significant text mining operations are Information Retrieval, Text Clustering, Text Categorization, Text searching and Text Summarization. Text mining predominantly used in different areas such as Record Management, Biomedical, Enterprise business Intelligent, Security applications, Social media and Marketing analysis. Information retrieval is one of the popular applications of text mining which helps to extract the relevant information based on the user requirements. In order to perform the information retrieval process the first and essential step is to classify the information available in the data sets. In this research work, information retrieval data set is used for performing classification. Classification tree algorithms are used for classifying the data set based on the attribute information category. The algorithms tested are Random forest algorithm, Random tree algorithm and Logistic model tree algorithm (LMT). Accuracy and execution time are the performance factors used in this comparative analysis. From the results, it is observed that the Random Forest algorithm efficiency is better than other algorithms.

Keywords: Data Mining, Information Retrieval, Tree classifier, Random Forest, Random Tree, Logistics Model Tree.

I. INTRODUCTION

Data mining can be defined as extraction of hidden information from the large database. Machine learning techniques are applied in the domain of Data Mining. Major data mining domains are text mining, sequence mining, spatial mining, image mining, structure mining, web mining, graph mining ect. Text Mining is the art of retrieving information from the group of unstructured data. Major roles of text mining are text categorization, text clustering and document summarization. Information Retrieval is a technology is used for extracting structured data and unstructured data. It supports the users to store, browse and retrieve the information. Consequently, Information retrieval poses several challenges for data mining algorithm. A major improvement in classification accuracy has resulted from growing an assembly of trees and letting them vote for the most admired class.

Potency of individual decision tree and correlation among base trees are key issues which decide generalization error of Tree classifiers. The Classification tree algorithms used and tested in this work are Random Forest, Random Tree and Logistic Model Tree algorithms. Comparative analysis is done by using Waikato Environment for Knowledge Analysis or in short, WEKA. This paper reviews and analyzes the basic concepts, applications, various algorithms and techniques used in information retrieval domain.

II. Literature Review

Albert A.Montillo[2] have analysed about the Random forest and their randomization methods. He also made the comparison between the Random forest and Boosting. He implemented an algorithm and flow chart design for the working of Random Forest.

Aaditya Desai, et.al [1] conducted an experimental study of real world problems using the WEKA implementations of Machine Learning algorithms and mainly perform classification and comparison of relative performance of different algorithms under certain criteria.

Vrushali Y Kulkarni, et.al, [13] analysed an in-depth study related to Random Forest that help to accelerate research in the field of Machine Learning. They presented a systematic survey of work done in Random Forest area. In their process, they derived Taxonomy of Random Forest Classifier and also prepared a Comparison chart of existing Random Forest classifiers on the basis of relevant parameters.

D.L Gupta et.al[4] have analysed the four Classification Tree algorithms namely Random Forest, J48, Reduce Error Pruning Tree, Logistics Model Tree, from these algorithms they have concluded that the Random forest produces the best accuracy measure using the weather datasets

Niels Landwehr[11] have implemented a new algorithm for solving classification problems, called the logistic model tree learner, and weigh against it to other modern learning schemes on several real-world problems.

Sam Drazin, et.al.,[12] Analysed the testing data sets using a variety of open source machine learning algorithms. The data sets were tested using the J48 Decision tree inducing algorithm. And examined the Two pruning methods namely post pruning and online pruning method. The overlying techniques had compared the amount of error rate in decision trees.

III. Methodology

Information Retrieval is used to mine data from huge set of databases. Using Functions and Trees classification we find the best algorithm for effective Information Retrieval. The architecture of the comparative analysis is as follows:

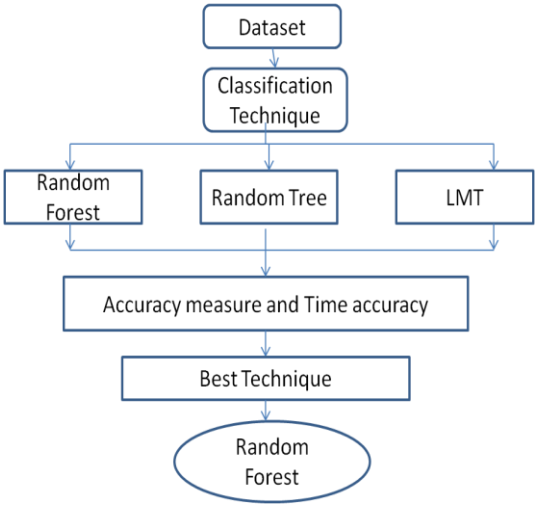


Fig 1: Flow Chart for Tree Classifiers

A. Dataset

The synthetic dataset have been created for analysis. This dataset contains 1K instances and 6 attributes. Weka data mining tool is used for analysing the performance of the classification algorithms.

B. Classification

Classification is a data mining practice used to foresee group membership for data instances. Classification is comparable to clustering in that it also segment Information Retrieval into distinct segment called classes. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. This paper analyses the three classifiers algorithms namely Random Forest, Random Tree, Logistics Model Tree using the accuracy measure, error rate and the time taken.

C. Tree Classifier

A tree is one which supports graph like representations from which we can get back the possible outcomes. A tree is a one which resembles like a flow chart structure where the internal node refers to the attributes and the leaf nodes represents the class label. A path from root to leaf represents the classification rules. Tree has three types of nodes namely Decision nodes, Chance nodes and

End nodes. Some of the advantages of the trees are they are simple to understand and implement. The process of the classification tree can be classified into the following steps: Selecting the criteria for predictive accuracy, selecting the spilt, determine the time to stop splits in a tree, and finally choosing the best tree. Best, average, worst cases can determine easily.

Random Forest: It is one of the most popular technique for tree classifier. It can be used for classification and regression. Random forest has a many number of predictors where each tree depends on the value of random vector. Random forest nurture many number of classification trees. The core principle is that the "strong learners" group is being formed by the group of "weak learners". The Random Forest algorithm was developed by Leo Brieman and Adele Cutler. Features of Random Forest are that they can handle a very large number of input variables, and it can run resourcefully on large databases, accurate results, fast and robust. A famous algorithm was used for Random Forest which was developed by Leo Breiman and Adele Cutler.

Algorithm:

Start
Step1: Let N ‹— denote the number of training cases
Sterp2: Let M ‹— denote the number of variables in classifier
/* choose a training set tree */
Step3: Select a new bootstrap model for the training sets.
Step4: In the bootstrap un-pruned trees must be developed.
Step5: Randomly select the predictors
Step6: Predict the best splits from the tree
End

Random Trees: It is a tree or a tree like structure that describes the possible outcomes through prediction. The classification works as follows: the random trees classifier takes the input vector, classifies it with every tree in the forest, and outputs the class labels. All the trees are trained with the same parameters. A random tree is a tree constructed randomly from a set of possible trees having K random features at each node. "At random" in this context means that in the set of trees each tree has an equal chance of being sampled. Or we can say that trees have a "uniform" distribution. Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models. There has been an extensive research in the recent years over. Random trees in the field of machine Learning. The types of the random trees include:

- **Uniform Spanning Tree:** It is a tree where each different tree is equally likely to be selected.
- **Random minimal spanning tree:** It is a tree which is formed by the minimum edge weights.
- **Random binary tree:** Is a tree which formed by inserting the nodes in the random order or by selecting all the possible trees.
- **Treap or Randomized binary search tree:** This uses random choices to stimulate a random binary tree.
- **Brownian Tree:** It is a fractal tree structure created by diffusion by limited aggregation processes.

Logistics Model Tree: A logistics model tree is the one which consists the decision tree structure with logistic regression functions. The logistic model tree has a tree structure consisting of inner or non-terminal nodes N and set of leave or terminal nodes T. Let $N = E_1 \times .... \times E_m$ denotes the space, spanned by all the attributes which are present in the data. LMT produces a single outcome in the form of tree containing binary splits on numeric attributes.

Algorithm of Logistics Model Tree:

The construction of the logistics model tree consists of the following scenario:

1. Growing the tree
2. Logistics Tree construction
3. Pruning tree

Growing the Tree:

Step1: Construct the tree using nominal and numeric values

Step2: Create a splitting criteria for ending the tree construction

Building the Logistics Model: The construction of the logistics model is based upon the attribute selection, for knowing what are all the attributes used and how the attributes are used. This also determines how to handle the numerical and nominal values while structuring a tree.

Pruning of the Tree: A fully grown Logistics model tree will cause lot of problems like over fitting in the dataset. So, Pruning is the art of cutting or removing the branches of the trees of the fitting the tree perfectly within the datasets

IV. Experimental Results

A. Accuracy Measure and Error Rate

Accuracy

The term accuracy refers to the correctly classified instances by the total number of instances present in the dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Where TP- True Positive, FP- False Positive, TN- True Negative, FN- False Negative

TP Rate: It is the ability which is used to find the high true-positive rate . The true-positive rate is also called as sensitivity.

$$TPR = \frac{TP}{TP + FN}$$

Precision

Precision is the ratio of modules correctly classified to the number of entire modules classified fault-prone. It is proportion of units correctly predicted as faulty.

$$Precision = \frac{TP}{TP + FP}$$

F-Measure

F- Measure is the one has the combination of both precision and recall which is used to compute the score. This kind of measure is frequently used in the field of Information Retrieval to calculate the query classification performance.

$$F - Measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

Confusion Matrix

The confusion matrix is used to measure the performance of two class problem for the given data set. The right diagonal elements TP (true positive) and TN (true negative) correctly classify Instances as well as FP (false positive) and FN (false negative) incorrectly classify Instances.

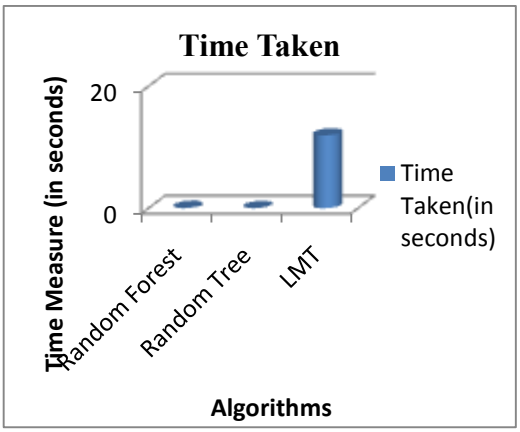


Fig 2: Time taken for building model

Above chart is used to represent the time taken by the tree classifiers. From the given chart we declare that Random Forest takes less amount of time to calculate the datasets. This chart represented as below given table 1.

Algorithm	Time taken to build
Random Forest	0.08 seconds
Random Tree	0.03 seconds
LMT	11.94

Table 1. Time taken for building model

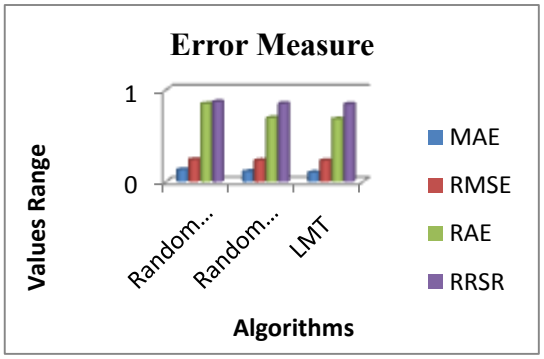


Fig 3: Error Measure for Tree Classifier

The above given chart calculates the accuracy error measure for tree classifier using the algorithms Random Forest, Random Tree and Logistic Model Tree. From the chart we declare that the RRSR of Random Forest act best for Function classifier. This chart represented as below given table2.

Algorithm	MAE	RMSE	RAE	RRSR
Random Forest	0.1342	0.245	84.571	87.04
Random Tree	0.11	0.23	69.29	84.93
LMT	0.108	0.237	68.099	84.23

Table 2: Error Measure for Tree Classifier

Table 3: Accuracy Measure for Tree Classifiers

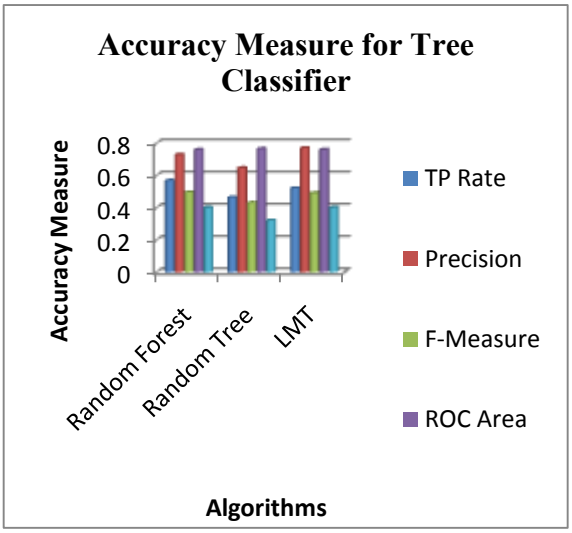


Fig 4: Accuracy Measure for Tree Classifier

The above given chart calculates the accuracy measure for tree classifier using the algorithms Random Forest, Random Tree and Logistic Model Tree. From the chart we declare that the Random Forest act best for Tree classifier according to correctly classified instances. This chart represented as below given table3.

Algorithm	Correctly classified Instances value (%)	Incorrectly classified Instances value (%)	TP Rate	Precision	F-measure	ROC Area	Kappa Statistics
Random Forest	**56.80**	**43.20**	**0.568**	**0.72**	**0.54**	**0.87**	**0.46**
Random Tree	**46.40**	**53.60**	**0.464**	**0.646**	**0.430**	**0.765**	**0.3258**
LMT	**52.40**	**47.60**	**0.524**	**0.776**	**0.495**	**0.760**	**0.4052**

V. Conclusions

Data Mining refers to extracting the useful information from the unstructured or structured data from in the large databases. In this paper we have classification algorithms to know the best classifier in the sections of Tree based algorithms. By analysing the algorithms we conclude that the Random Forest suits the best from the Tree Classifier algorithms.

REFERENCES

1. Aaditya Desai, et.al., "Analysis of Machine Learning Algorithms using WEKA " International Conference & Workshop on Recent Trends in Technology, (TCET) 2012 Proceedings published in International Journal of Computer Applications® (IJCA)
2. Albert A.Montillo "Random Forests" http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf
3. Data Mining Classification: Decision Trees http://staffwww.itn.liu.se/~aidvi/courses/06/dm/lectures/lec3.pdf
4. Gupta D. L et.al " Performance Analysis of Classification Tree Learning Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 55– No.6, October 2012
5. Jehad Ali et.al "Random Forests and Decision Trees" IJCSI International Journal of Computer Science and Issues, Vol 9, Issue 5, No.3, Sept2012.
6. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques" (2nd Edition), Morgan Kaufmann Publisher, (2006)
7. Kaushik H. Raviya et al., "Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA" ISSN - 2250-1991 Volume : 2 | Issue : 1 | January 2013
8. Leo Breiman "Random Forests" http://oz.berkeley.edu/~breiman/randomforest2001.pdf
9. Marc Summner et.al "Speeding up Logistics Model Tree Induction " www.cs.waikato.ac.nz/
10. Nathen Pater "Enhancing Random Forest Implementation in Weka" Machine learning Conference Paper for ECE591Q, Nov 28,2005
11. Niels Landwehr "Logistic Model Trees" Diploma of Computer Science at the University of Freiburg, Freiburg, Germany. 9th of July, 2003
12. Sam Drazin, et.al, "Decision Tree Analysis Using Weka" University of Miami.
13. Trilok Chand Sharma, et.al., "WEKA Approach for Comparative Study of Classification Algorithm" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013
14. Vrushali Y Kulkarni, etal., "Efficient Learning of Random Forest Classifier using Disjoint Partitioning Approach" Proceedings of the World Congress on Engineering 2013 Vol II, WCE 2013, July 3 - 5, 2013, London, U.K.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Survey of Various Security attacks and it issues in MANET

R.Rajeshkanna, P.Dineshkumar, S.Poorana Senthilkumar,

[1 2 3]Assistant Professor

Dept.of Computer Applications,

Dr.N G P Arts and Science Collgege, Coimbatore

ABSTRACT

Mobile ad-hoc networks can turn the dream of getting connected "anywhere and at any time" into reality. Typical application examples include a disaster recovery or a military operation. Not bound to specific situations, these networks may equally show better performance in other places. As an example, we can imagine a group of peoples with laptops, in a business meeting at a place where no network services is present. They can easily network their machines by forming an ad-hoc network. This is one of the many examples where these networks may possibly be used. The strength of the connection can change rapidly in time or even disappear completely. Nodes can appear, disappear and re-appear as the time goes on and all the time the network connections should work between the nodes that are part of it. As one can easily imagine, the situation in ad hoc networks with respect to ensuring connectivity and robustness is much more demanding than in the wired case.

Keywords: Routing, Security, Mobile Ad Hoc Networks

Introduction

Wireless cellular systems have been in use since 1980s. We have seen their evolutions to first, second and third generation's wireless systems. Wireless systems operate with the aid of a centralized supporting structure such as an access point. These access points assist the wireless users to keep connected with the wireless system, when they roam from one place to the other.

The presence of a fixed supporting structure limits the adaptability of wireless systems. In other words, the technology cannot work effectively in places where there is no fixed infrastructure. Future generation wireless systems will require easy and quick deployment of wireless networks. This quick network deployment is not possible with the existing structure of current wireless systems.

Recent advancements such as Bluetooth introduced a new type of wireless systems known as mobile ad-hoc networks. Mobile ad-hoc networks or "short live" networks operate in the absence of fixed infrastructure. They offer quick and easy network deployment in situations where it is not possible otherwise. Ad-hoc is a Latin word, which means "for this or for this only." Mobile ad-hoc network is an autonomous system of mobile nodes connected by wireless links; each node operates as an end system and a router for all other nodes in the network.

Nodes in mobile ad-hoc network are free to move and organize themselves in an arbitrary fashion. Each user is free to roam about while communication with others. The path between each pair of the users may have multiple links and the radio between them can be heterogeneous. This allows an association of various links to be a part of the same network.

The popular IEEE 802.11 "WI-FI" protocol is capable of providing ad-hoc network facilities at low level, when no access point is available. However in this case, the nodes are limited to send and receive information but do not route anything across the network. Mobile ad-hoc networks can operate in a standalone fashion or could possibly be connected to a larger network such as the Internet.

Mobile ad-hoc networks can turn the dream of getting connected "anywhere and at any time" into reality. Typical application examples include a disaster recovery or a military operation. Not

bound to specific situations, these networks may equally show better performance in other places. As an example, we can imagine a group of peoples with laptops, in a business meeting at a place where no network services is present. They can easily network their machines by forming an ad-hoc network. This is one of the many examples where these networks may possibly be used.

Characteristics

Mobile Adhoc Network (MANET) is a collection of independent mobile nodes that can communicate to each other via radio waves. The mobile nodes that are in radio range of each other can directly communicate, whereas others needs the aid of intermediate nodes to route their packets. These networks are fully distributed, and can work at any place without the help of any infrastructure.

The characteristics of these networks are summarized as follows:

• Communication via wireless means.

• Nodes can perform the roles of both hosts and routers.

• No centralized controller and infrastructure. Intrinsic mutual trust.

• Dynamic network topology. Frequent routing updates.

• Autonomous, no infrastructure needed.

• Can be set up anywhere.

• Energy constraints

• Limited security

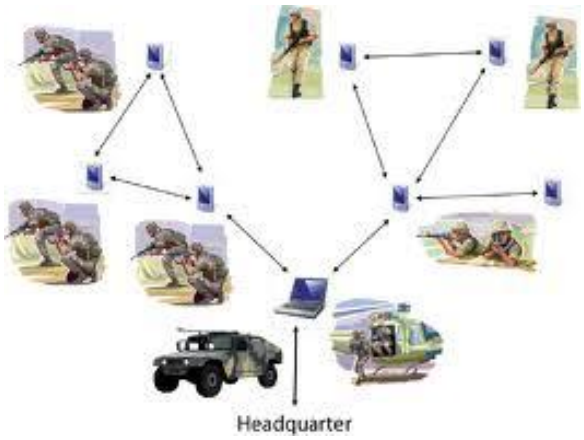


Fig.1 Mobile Adhoc Network

The dynamical nature of the network topology increases the challenges of the design of ad hoc networks. Each radio terminal is usually powered by energy limited power source (as rechargeable batteries). The power consumption of each radio terminal could be divided generally into three parts, power consumption for data processing inside the RT, power consumption to transmit its own information to the destination, and finally the power consumption when the RT is used as a router, i.e. forwarding the information to another RT in the network. The energy consumption is a critical issue in the design of the ad hoc networks.

The mobile devices usually have limited storage and low computational capabilities. They heavily depend on other hosts and resources for data access and information processing. A reliable network topology must be assured through efficient and secure routing protocols for Ad Hoc networks.

Application Areas

Some of the applications of MANETs are

• Military or police exercises.

• Disaster relief operations.

• Mine site operations.

• Urgent Business meetings

• Robot data acquisition

It is easy to imagine a number of applications where this type of properties would bring benefits. One interesting research area is inter-vehicle communications. It is one area where the ad hoc networks could really change the way we communicate covering personal vehicles as well as professional mobile communication needs. Also, it is area where no conventional (i.e. wired) solutions would do because of the high level of mobility. When considering demanding surroundings, say mines for example, then neither would the base station approach work but we must be able to accomplish routing via nodes that are part of the network i.e. we have to use ad hoc network.

Such networks can be used to enable next generation of battlefield applications envisioned by the military including situation awareness systems for maneuvering war fighters, and remotely deployed unmanned micro-sensor networks. Ad Hoc networks can provide communication for civilian applications, such as disaster recovery and message exchanges among medical and security personnel involved in rescue missions.

Routing in MANETs

Routing in mobile ad hoc networks faces additional problems and challenges when compared to routing in traditional wired networks with fixed infrastructure. There are several well-known protocols in the literature that have been specifically developed to cope with the limitations imposed by ad hoc networking environments. The problem of routing in such environments is aggravated by limiting factors such as rapidly changing topologies, high power consumption, low bandwidth and high error rates . Most of the existing routing protocols follow two different design approaches to confront the inherent characteristics of ad hoc networks, namely the table-driven and the source-initiated on-demand approaches. The following sections analyze in more detail these two design approaches, and briefly present example protocols that are based on them.

Table-driven Ad hoc Routing Protocols

Table-driven ad hoc routing protocols maintain at all times routing information regarding the connectivity of every node to all other nodes that participate in the network. Also known as proactive, these protocols allow every node to have a clear and consistent view of the network topology by propagating periodic updates . Therefore, all nodes are able to make
immediate decisions regarding the forwarding of a specific packet. On the other hand, the use of periodic routing messages has the effect of having a constant amount of signaling traffic in the network, totally independent of the actual data traffic and the topology changes. As an example of two protocols that follow the table-driven design approach we will briefly present the Destination-Sequenced Distance-Vector (DSDV) protocol and the Optimized Link State Routing (OLSR) protocol
.

Destination-Sequenced Distance-Vector Routing (DSDV)

DSDV is a table-driven routing protocol based on the Bellman-Ford algorithm. The DSDV protocol can be used in mobile ad hoc networking environments by assuming that each participating node acts as a router. Each node must maintain a table that consists of all the possible destinations. In more detail, an entry of the table contains the address identifier of a destination, the shortest known distance metric to that destination measured in hop counts and the address identifier of the node that is the first hop on the shortest path to the destination. Furthermore, the DSDV protocol adds a sequence number to each table entry assigned by the destination node, preventing the formation of routing loops caused by stale routes. The routing tables are maintained by periodically transmitted updates by each router to all the neighboring routers.

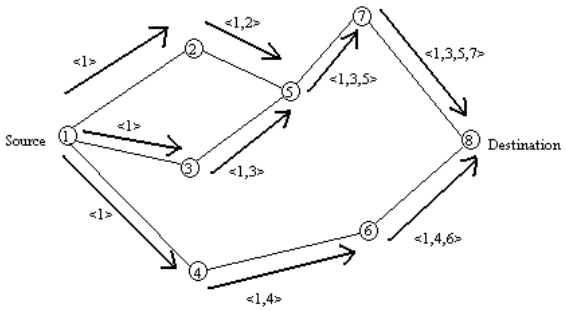
Optimized Link State Routing (OLSR)

The Optimized Link State Routing (OLSR) protocol is a proactive link state routing protocol based on the Open Shortest Path First (OSPF) protocol. OLSR has been specifically developed to support mobile ad hoc networks and the constraints they impose on routing. The OLSR protocol can be conceptually divided into three different operations, namely neighbor sensing, distribution of signaling traffic and distribution of topological

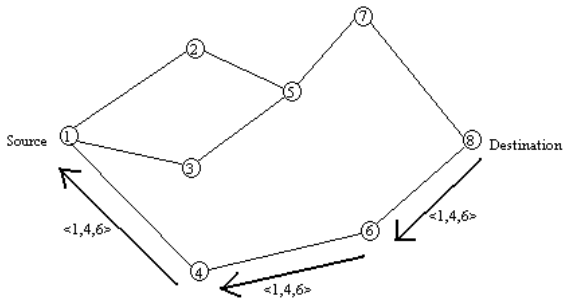
information . Neighbor sensing in OLSR is accomplished by transmitting periodic hello messages that contain the generating node's address identifier, a list of its neighboring nodes and the type of the link it has with each neighbor (e.g.: symmetric or asymmetric). For the distribution of signaling traffic OLSR adopts a flooding mechanism where every node forwards a flooded message that it has not forwarded previously. Finally, the distribution of topological information function is realized with the use of periodic topology control messages that result in each node knowing a partial topology graph of the network which is then used for the computation of optimal routes.

Source-initiated On-demand Ad hoc Routing Protocols

An alternative approach to the one followed by table-driven protocols is the source initiated on-demand routing. According to this approach a route is created only when the source node requires one to a specific destination. A route is acquired by the initiation of a route discovery function by the source node. The data packets transmitted while a route discovery is in process are buffered and are sent when the path is established. An established route is maintained as long as it is required through a route maintenance procedure. The Ad hoc On-demand Distance Vector (AODV) routing protocol and the Dynamic Source Routing protocol are examples of this category of protocols also known as reactive



(a) Building Record Route during Route Discovery



(b) Propogation of Route Reply with the Route Record

Ad hoc On-demand Distance Vector Routing (AODV)

The AODV protocol uses route request (RREQ) messages flooded through the network in order to discover the paths required by a source node. An intermediate node that receives a RREQ replies to it using a route reply message only if it has a route to the destination whose corresponding destination sequence number is greater or equal to the one contained in the RREQ. This effectively means that an intermediate node replies to a RREQ only if it has a fresh enough route to the destination. Otherwise, an intermediate node broadcasts the RREQ packet to its neighbors until it reaches the destination. The destination unicast a RREP, back to the node that initiated the route discovery by transmitting it, to the neighbor from which it received the RREQ.

Dynamic Source Routing (DSR)

The Dynamic Source Routing (DSR) protocol is based on a method known as source routing . The route discovery process in DSR is similar to

the one used by AODV, except that each intermediate node that broadcasts a route request packet adds its own address identifier to a list carried in the packet. The destination node generates a route reply message that includes the list of addresses received in the route request and transmits it back along this path to the source. Route maintenance in DSR is accomplished through the confirmations that nodes generate when they can verify that the next node successfully received a packet. These confirmations can be link-layer acknowledgements, passive acknowledgements or network-layer acknowledgements specified by the DSR protocol. When a node is not able to verify the successful reception of a packet it tries to retransmit it. When a finite number of retransmissions fail, the node generates a route error message that specifies the problematic link, transmitting it to the source node.

Security Problem with Existing Ad hoc Routing Protocols

The main assumption of the previously presented ad hoc routing protocols is that all participating nodes do so in good faith and without maliciously disrupting the operation of the protocol. However, the existence of malicious entities cannot be disregarded in any system, especially in open ones like ad hoc networks. In ad hoc network the routing function can be disrupted by internal or external attackers. An internal attacker can be any legitimate participant of the routing protocol. An external attacker is defined as any other entity. Cryptographic solutions can be employed to prevent the impact of external attackers by mutual authentication of the participating nodes through digital signature schemes. However, the underlying protocols should also be considered since an attacker could manipulate a lower level protocol to interrupt a security mechanism in a higher level. Internal attackers having capability to complete access the communication link they are able to advertise false routing information at will and force arbitrary routing decisions on their peers .

Security Goals

In providing a secure networking environment some or all of the following service may be required

Authentication: This service verifies the identity of node or a user, and to be able to prevent impersonation. In wired networks and infrastructure-based wireless networks, it is possible to implement a central authority at a point such as a router, base station, or access point. But there is no central authority in MANET, and it is much more difficult to authenticate an entity. Authentication can be providing using encryption along with cryptographic hash function, digital signature and certificates.

Confidentially: Keep the information sent unreadable to unauthorized users or nodes. MANET uses an open medium, so usually all nodes within the direct transmission range can obtain the data. One way to keep information confidential is to encrypt the data, and another technique is to use directional antennas. It also ensures that the transmitted data can only be accessed by the intended receivers.

Integrity: Ensure that the data has been not altered during transmission. The integrity service can be provided using cryptography hash function along with some form of encryption. When dealing with network security the integrity service is often provided implicitly by the authentication service.

Availability: Ensure that the intended network security services listed above are available to the intended parties when required. The availability is

typically endure by redundancy, physical protection and other non-cryptographic means, e.g. use of robust protocol.

Non-repudiation: Ensure that parties can prove the transmission or reception of information by another party, i.e. a party cannot falsely deny having received or sent certain data. By producing a signature for the message, the entity cannot later deny the message. In public key cryptography, a node A signs the message using its private key. All other nodes can verify the signed message by using A's public key, and A cannot deny that its signature is attached to the message.

Access Control: To prevent unauthorized use of network services and system resources. Obviously, access control is tied to authentication attributes. In general, access control is the most commonly thought of service in both network communications and individual computer systems.

Vulnerability in MANETs

Malicious and selfish nodes are the ones that fabricate attacks against physical, link, network, and application-layer functionality. Current routing protocols are exposed to two types of attacks:

- Active attacks
- Passive attacks

Classification of Security Attacks

Active attacks	Spoofing, Fabrication, Wormhole Attack, Modifi cation, Denial of Service
Passive Attacks	Eavesdropping, traffic analysis, monitoring

Active Attacks

Active attacks are the attacks that are performed by the malicious nodes that bear some energy cost in order to perform the attacks. Active attacks involve some modification of data stream or creation of false stream. These attacks can be classified into further following types.

Spoofing: Occurs when a malicious node misrepresents its identity in order to alter the vision of the network topology that a benign node can gather .

Fabrication: The notation "fabrication" is used when referring to attacks performed by generating false routing messages. Such kind of attacks can be difficult to identify as they come as valid routing constructs, especially in the case of fabricated routing error messages, which claim that a neighbor can no longer be contacted.

Wormhole Attack: An attacker records packets at one location in the network and tunnels them to another location. Routing can be disrupted when routing control messages are tunneled. This tunnel between two colluding attackers is referred as a wormhole. Wormhole attacks are severe threats to MANET routing protocols.

Modification: The attacker performs such attacks is targeted to integrity of data, by altering packet or modifying packets.

Denial of Service: This active attack aims at obstructing or limiting access to a certain resource. The resource can be a specific node or service or the whole network. The nature of ad-hoc networks, where several routes exist between nodes and routes are very dynamic gives ad hoc a built-in resistance to Denial of Service attacks, compared to fixed networks.

Passive Attacks

In passive attacks the attacker does not perturb the routing protocol, instead try to extract the valuable

information like node hierarchy and network topology from it. Passive attack is in nature of eavesdropping on, or monitoring of, transmission. The goal of opponent is to obtained information that is being transmitted . Passive attacks are very difficult to detect because they do not involve any alteration of data.

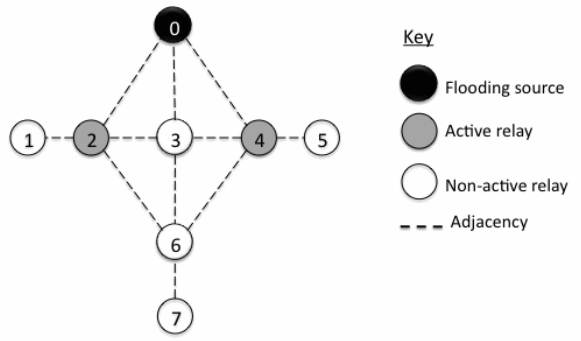
Other Advanced Attacks

We will now discuss several specific attacks that can affect the operation of a routing protocol in ad hoc network.

Blackhole attack: In a black hole attack a malicious node advertising itself as having a valid route to the destination. With this intension the attacker consume or intercept the packet without any forwarding . An attacker can completely modify the packet and generate fake information, this cause the network traffic diverted or dropped.

Byzantine attack: A compromised with set of intermediate, or intermediate nodes that working alone within the network carry out attacks such as creating routing loops, forwarding packets through non-optimal paths, or selectively dropping packets, which results in disruption or degradation of the routing services within the network .

Rushing attack: Two colluded attackers use the tunnel procedure to form a wormhole. If a fast transmission path (e.g. a dedicated channel shared by attackers) exists between the two ends of the wormhole, the tunneled packets can propagate faster than those through a normal multi-hop route. This forms the rushing attack . The rushing attack can act as an effective denial-of-service attack against all currently proposed on-demand MANET routing protocols, including protocols that were designed to be secure, such as ARAN and Ariadne .



Replay attack: An attacker that performs a replay attack are retransmitted the valid data repeatedly to inject the network routing traffic that has been captured previously. This attack usually targets the freshness of routes, but can also be used to undermine poorly designed security solutions.

Location disclosure attack: An attacker discover the Location of a node or structure of entire networks and disclose the privacy requirement of network through the use of traffic analysis techniques, or with simpler probing and monitoring approaches. Adversaries try to figure out the identities of communication parties and analyze traffic to learn the network traffic pattern and track changes in the traffic pattern. The leakage of such information is devastating in security.

Secure Ad hoc Routing

There exist several proposal that attempt to architect a secure routing protocol for mobile ad hoc network , in order to offer protection against the attacks mentioned in the previous section. There are several solutions proposed by researcher they are either completely new stand-alone protocol or in some cases incorporation of security mechanism into existing one like DSDV and AODV . Since routing is an essential function for ad hoc networks, the integrated security procedures should not hinder its operation. Another important part of analysis is the examination of assumption and the requirements that each solution depend on.

Although a protocol might be able to satisfy certain security constraints, its operational requirements might thwart its successful employment. In order to analyze exiting solution in structure way we have classified them into three categories; Solution based on Symmetric cryptography, solution based on Asymmetric cryptography and Hybrid solution. However, this classification is only indicative since a lot of solution can be classified into more than one category.

Symmetric Cryptography Solutions

The Secure Efficient Ad hoc Distance Vector (SEAD) is a secure ad hoc network routing protocol based on the design of the Destination-Sequenced Distance-Vector (DSDV) algorithms. To developing SEAD, follow the table driven approach. In table driven routing protocol maintain at all times routing information regarding to the network connectivity of every node to all other nodes. It is also known as proactive routing protocol. In order to find shortest path between two nodes, the distance vector routing protocol utilize a distributed version of Bellman Ford Algorith. The SEAD routing protocol employ the use of hash chains to authenticate hop count and sequence numbers.

Network Security

Network security extends computer security, thus all the things in computer security are still valid, but there are other things to consider as well. Computer security is defined as follows:

Broadly speaking, security is keeping anyone from doing things you do not want them to do to, with, or from your computers or any peripherals- - William R. Cheswick.

Network security is then computer security plus secure communication between the computers or other devices. Not all nodes are computers in an Ad Hoc network, thus nodes cannot be assumed to implement the security services normally existent in computers' operating systems. That is why network security should be defined as:

-Making sure that the nodes enforce a proper computer security and then securing the communication between them-.

Different variables have different impact on security issues and design. Especially environment, origin, range, quality of service and security criticality are variables that affect the security in the network.

If the environment is concerned, networks can operate in hostile or friendly environments. A battlefield has totally different requirements for security if compared with home networks. On a battlefield also physical security and durability might be needed to ensure the functionality of the network.

The ways to implement security vary if the range of the network varies. If the nodes are very far from each others, the risk of security attacks increases. On the other hand, if the nodes are so close to each others that they actually can have a physical contact, some secret information (e.g. secret keys) can be transmitted between the nodes without sending them on air. That would increase the level of security, because the physical communication lines are more secure than wireless communication lines.

Quality of service issues deal with questions like - Is it crucial that all messages reach their destination?- or -Is it crucial that some information reaches the destination in certain time?-. QoS is generally demanded in real time applications where reliable and deterministic communication is required. These issues refer to security e.g. in process control applications. We could have an Ad Hoc network in some process and all the measurements and control signals could be transmitted through the network. In order to have secure and reliable control of the process, quality of service requirements need to be met.

The last variable of Ad Hoc networks described with respect to security is security criticality. This means that before we think of the ways to implement security, we must consider carefully whether security is required at all or whether it matters or not if someone outside can see what packets are sent and what they contain. Is the network threatened if false packets are inserted and old packets are retransmitted? Security issues are not always critical, but it might cost a lot to ensure it. Sometimes there is trade-off between security and costs.

Security Problems in MANETs

MANETs are much more vulnerable to attack than wired network. This is because of the following reasons :

Open Medium - Eavesdropping is more easier than in wired network.

Dynamically Changing Network Topology - Mobile Nodes comes and goes from the network, thereby allowing any malicious node to join the network without being detected.

Cooperative Algorithms - The routing algorithm of MANETs requires mutual trust between nodes which violates the principles of Network Security.

Lack of Centralized Monitoring - Absence of any centralized infrastructure prohibits any monitoring agent in the system.

Lack of Clear Line of Defense - Experience of security research in wired world has taught us that we need to deploy layered security mechanisms because security is a process that is as secure as its weakest link.

Advantages

The following are the advantages of MANETs:

• They provide access to information and services regardless of geographic position.

• These networks can be set up at any place and time.

• These networks work without any pre-existing infrastructure.

Disadvantages

Some of the disadvantages of MANETs are:

• Limited resources. Limited physical security.

• Intrinsic mutual trust vulnerable to attacks. Lack of authorization facilities.

• Volatile network topology makes it hard to detect malicious nodes.

Security protocols for wired networks cannot work for ad hoc networks.

Conclusion

Ad hoc networks can be implemented using various techniques like Bluetooth or WLAN for example. The definition itself does not imply any restrictions to the implementing devices.

Ad Hoc networks need very specialized security methods. There is no approach fitting all networks, because the nodes can be any devices. The computer security in the nodes depends on the type of node, and no assumptions on security can be made. In this paper the computer security issues have not been discussed, because the emphasis has been on network security.

But with the current MAC layer and routing solutions the true and working ad hoc network is just a dream for now. However, it can be used with relatively small networks and potentially some very nice applications can be realized. Although some peer-to-peer type of solutions work nicely already today, it would be nice to see that some new and innovative solutions would be seen in the arena of ad hoc networks since it is not hard for one to imagine a countless number of nice ad hoc based applications that would make the world at least a bit better place.

References

[1] C.-K Toh, Ad Hoc Mobile Wireless

Networks: Protocols and Systems, Prentice Hall, New Jersey, pp: 34-37, 2007.

[2] Larry Stotts, Scott Seidel, Tim Krout, and Paul Kolodzy, "MANET Gateways: Radio Interoperability via the Internet, Not the Radio",
IEEE Communications Magazine, June 2008, Vol 46. No. 6, PP.51-60.

[3] Lt Col R J B Spencer, and Jim Ironside, "Network Centric Warfare Operation in an Expeditionary Context", Military Information & communications, Symposium of South Africa (MICSSA), July 2007.

[4] "Trust Management System for Network of Networks
(NoN)",http://www.cse.dmu.ac.uk/STRL/research/nato/index.html, visited June 28, 2008.

[5] C. Siva Ram Murthy, and B.S. Manoj, Ad Hoc Wireless Networks: Architectures and Protocols, Prentice Hall communications engineering and emerging technologies series Upper Saddle River, New Jersey, 2004.

[6] Mohammed Ilyas, The Handbook of Ad Hoc Wireless Networks, CRC Press, New York, 2003.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Applications of spatial data mining using GPS Technology

Mrs, D.Gandhimathi M.C.A.,M.Phil. [2]**M.Ramya priya,**
[1]Assistant Professor, BCA & SS Department,
Sri Krishna Arts and Science College,Coimbatore.

ABSTRACT

A growing attention has been paid to spatial data mining and knowledge discovery. Spatial data mining is to extract implicit knowledge, spatial relations as well as other non-explicitly information which included in the spatial database but it exist with other model from the spatial database. Due to the complexity of spatial data and its application of the professional, and it gives challenges to spatial data mining. Spatial data acquisition is one of the most expensive and time- consuming tasks in the deployment and Updating of Geographic Information Systems (GIS). Global positioning system is an artificial satellite system which plays a major application in tracking system, navigation, aeronautics, etc. This paper discusses about spatial data mining architecture, Global positioning system, its general applications and its working in automobiles are discussed below

Key terms:

Spatial data mining, global positioning system, geographic information system

1. Introduction:

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information .data processing using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large preexisting databases. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets [1]. The Global Positioning System (GPS) is a satellite-based navigation system made up of a network of 24 satellites placed into orbit by the U.S. Department of Defense.GIS is a relatively broad term that can refer to a number of different technologies, processes, and methods. is attached to many operations and has many applications related to engineering planning, management, insurance, telecommunications, and business. For that reason, GIS and location intelligence applications can be the foundation for many location-enabled services that rely on analysis, visualization and dissemination of results for collaborative decision making.

2. Spatial data mining:

Spatial data mining is the application of data mining methods to spatial data. The end objective of spatial data mining is to find patterns in data with respect to geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions, and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasize the importance of developing data-driven inductive approaches to geographical analysis and modeling [1].

3. Spatial data mining architecture:

Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Figure 1illustrate architecture for advanced analysis in a large data warehouse.

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure [2].

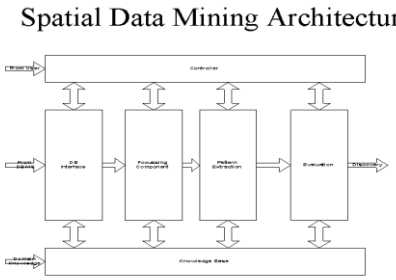


Fig 1: Spatial data mining architecture

4. Global positioning System:

GPS (Global Positioning System) was an artificial satellite system being composed of 24 artificial satellites covering entire earth, as shown in figure 2. The system can guarantee that at any moment, any point on earth, and at least four satellites can be observed contemporaneously, to ensure that the satellites can be collected to the latitude, longitude and height of the observation points, in order to achieve navigation, positioning and timing capabilities. GPS global positioning system was consisted of three parts: the space part - GPS constellation; ground control part - the ground monitoring system; users of the equipment - GPS signal receivers. GPS positioning technology with the advantages of all-weather, high coverage, high precision, high efficiency and low cost, has been more wide range of application in all types of geodetic control network to strengthen reform and establishment, and in highway engineering survey and large-scale structures of measuring deformation measurement.

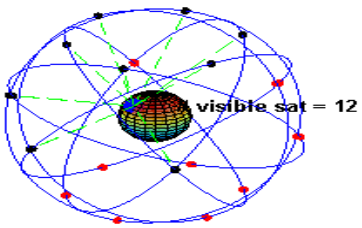


Fig 2: Global positioning system with 24 artificial satellites

GPS positioning including absolute positioning and relative positioning: Absolute positioning methods: absolute positioning also called single-point position, it refers to directly determine the coordinates of observation stations with the coordinate system whose center was the Earth's center as coordinate origin. The principle was to take distance between consumer receiver air wire and GPS artificial satellite observation amounts as basis, and according to the known instantaneous satellite coordinates, to determine the user receiver antenna at the corresponding point coordinates. The micro time satellite bell and the receiver bell were difficult to keep strict synchronism, so the distance between the actual observation points and the satellite contains the distance affected by bell difference; this distance was called "pseudo-range ". The satellite bell error can be corrected by bell difference parameters of navigation message, but the receiver clock error cannot know in advance, so it ought to go to find the solution together as a unknown with the three-dimensional coordinates of observation points in the data processing, so a Real-time observation point to solving four unknowns using at least four satellites which were observed at the same time [3].

5. GPS working:

A GPS is basically divided into 3 major components as shown in Figure 3.These are the user segment, the control segment, and the space segment.

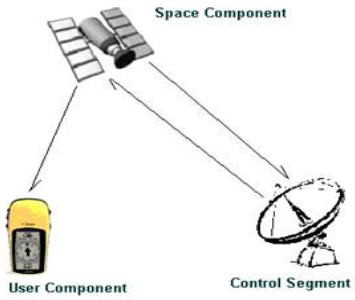


Fig 3: components of GPS

The space segment is composed of the GPS satellites1 that transmit time and position in the form of radio signals to the user. The whole set of 24 satellites is called a constellation. The control segment is composed of all the ground-based facilities that are used to monitor and control the satellites. This segment is usually not observed by the user. The user segment consists of the users and GPS receivers. A GPS receiver is a specialized radio receiver designed to listen to the radio signals being transmitted from the satellites.[6]

6. General Applications of GPS:

Surveying and Mapping, on land, at sea and from the air. The applications are of relatively high accuracy, for positioning in both the stationary and moving mode. Includes geophysical and resource surveys, GIS data capture surveys, etc.

Land, Sea and Air Navigation, including enroute as well as precision navigation, cargo monitoring, vehicle tracking, etc

Search and Rescue Operations, including collision avoidance and rendezvous functions.

Space craft Operations

Military Applications

Recreational Uses, on land, at sea and in the air other specialized uses, such as time transfer, attitude determination, automatic operation, etc [4]

7. GPS in cars:

The automobile industry is the largest user of GPS technology and uses it in the following ways. The most well-known commercial use of GPS is in car navigation systems. Before commencing a trip, a car driver just needs to feed in his destination in the GPS receiver's screen. The car GPS navigation system makes you find your way easily and quickly.

A car navigation system combines GPS technology with a computer communication system to track a driver's progress on a digital map. It provides services like the shortest route, fastest route, etc. on roadways, especially on motorways, which, on daily basis, are hardly free of congestion. This application is best termed optimal car routing and is the most widely used commercial application of GPS technology. Major companies like Motorola, Tom-tom, Nokia, Philips, Sony, etc. dominate this market[7].



Fig 4: GPS technology in cars

The control of the Positioning System consists of different tracking stations that are located across the globe. These monitoring stations help in tracking signals from the GPS satellites that are continuously orbiting the earth. Space vehicles transmit microwave carrier signals. The users of Global Positioning Systems have GPS receivers that convert these satellite signals so that one can estimate the actual position, velocity and time.

The operation of the system is based on a simple mathematical principle called trilateration. Trilateration falls into two categories: 2-D Trilateration and 3-D Trilateration. In order to make the simple mathematical calculation the GPS receiver must know two things. First it must know the location of the place is to be traced by at least three satellites above the place. Second, it must know the distance between the place and each of those Space Vehicles. Units that have multiple receivers that pick up signals from several GPS satellites at a same time. These radio waves are electromagnetic energy that travels at the speed of light.

8. Tracking:

A GPS tracking system can work in various ways Figure 5 shows the tracking of cars.[5] From a commercial perspective, GPS devices are generally used to record the position of vehicles as they make their journeys. Some systems will store the data within the GPS tracking system itself (known as passive tracking) and some send the information to a centralized database or system via a modem within the GPS system unit on a regular basis (known as active tracking) or 2-Way GPS.

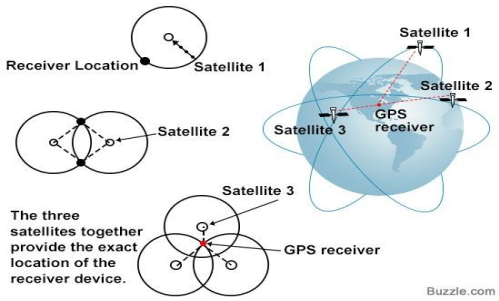


Fig 5: GPS tracking

A passive GPS tracking system will monitor location and will store its data on journeys based on certain types of events. So, for example, this kind of GPS system may log data such as where the device has traveled in the past 12 hours. The data stored on this kind of GPS tracking system is usually stored in internal memory or on a memory card, which can then be downloaded to a computer at a later date for analysis. In some cases the data can be sent automatically for wireless download at predetermined points/times or can be requested at specific points during the journey.

An active GPS tracking system is also known as a real-time system as this method automatically sends the information on the GPS system to a central tracking portal or system in real-time as it happens. This kind of system is usually a better option for commercial purposes such as fleet tracking or monitoring of people, such as children or elderly, as it allows a caregiver to know exactly where loved ones are, whether they are on time and whether they are where they are supposed to be during a journey. This is also a useful way of monitoring the behavior of employees as they carry out their work and of streamlining internal processes and procedures for delivery fleets.

9. GPS in spatial data mining:

The Global Positioning System as giving every centimeter on the earth a unique address. There are many different formats for displaying GPS coordinates. We adopt the floating point number system, otherwise referred to as the Decimal Format System, which ranges from -90 to 90 latitude and -180 to 180 longitudes [9].



Fig 6: spatial computing

Spatial geodetic technology opened up a brand new way for the current monitoring crustal movement, especially the GPS technology, with its observation of high precision, small size equipment, operating efficiency fast, not-as between stations, all-weather advantages etc., has brought revolutionary change for monitoring the global and regional crustal movement. Figure 6 shows about spatial computing, Using GPS to study crustal movement and explore geodynamic phenomenon in the world were developing rapidly, the application of GPS in geodesy, geophysics, geology, and other field, was more extensive, and the application of GPS observing technology were increasingly learning of the broad masses of workers to the attention[10].

10. Conclusion:

In this paper we have discussed about spatial data mining in GPS technology, general applications, role of GPS in car, and its tracking technology. Applications of spatial data mining play a vital role in GPS technology. “Acquiring preemptive knowledge about emerging technologies is the best way to ensure that we have a say in the making of our future.”

Reference:

[1].Krzysztof koperski, junas adhikary, jiawei han,”spatial data mining: progress and challenges survey paper” Simon Francis university, Canada,2000

[2].martin ester, Hans-peter kriegejorg sander,”algorithm and applications for spatial data mining” ,geographical data mining and knowledge discovery, research monographic in GIS ,francis,2001

[3].http://en.wikipedia.org/wiki/Global_Positioning_System

[4].http://www8.garmin.com/aboutGPS/applications.html

[5].http://www.buzzle.com/articles/how-location-tracking-works.html

[6]. stefanschroedl, kiri wagstaff, seth rogerls,pat Langley, Christopher willson,”mining GPS traces for map refinement”, data mining knowledge discovery, Netherlands,2004.

[7]. http://thecustomizewindows.com/2012/04/car-gps-how-a-car-gps-device-works/

[8].David luper, Detroy Cameron, john. A.miller,hamid.R.arbnid,”spatial and temporal target association through semantic analysis and GPS data mining”

[9].http://www.researchgate.net/publication/221333511_The_study_on_the_exploratory_spatial_data_mining_method_based_on_partial_random_walk_and_its_application_in_GPS_TEC_analysis[

10]. F.Limra, M.ferreire,”mining spatial data from GPS traces for automatic road networking extraction” in 6th international symposium on mobile mapping technology, soa Paulo, brazil, july 2000

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Heterogeneous MANET Using Multiclass Routing and Media Access Control

[1]V.Gayathri M.Sc.,(M.Phil) Research Scholar, [2]A.Vijaya M.Sc. M.Phil.,

[1 & 2] Sri Krishna Arts and Science College. Coimbatore.

ABSTRACT

In this paper mobile ad hoc network (MANETs) is very important for efficient routing model. Heterogeneous ad hoc networks consider existing routing protocols, in which all nodes are identical. The same communication capabilities and characteristics. A Mobile Ad-hoc network is a wireless ad-hoc network which is used to exchange information. Each node is willing to forward data to other nodes.It does not rely on fixed infrastructure.A heterogeneous network is a network connecting computers and other devices with different Operating systems or protocols.The heterogeneous network is also used in wireless networks using different access technologies. In this paper, present a new routing protocol called multiclass (MC) routing. It is specifically designed for heterogeneous MANETs. Heterogeneous MANETs is design for new medium access control (MAC) protocol, which is more efficient than IEEE 802.11b. In terms of reliability, scalability, route discovery latency, overhead, as well as packet delay and throughput.

I.INTRODUCTION

Mobile ad hoc networks are multi-hop wireless networks that are established by a group of mobile nodes on shared wireless channels. It is characterized by no fixed infrastructure, dynamic topologies, variable capacity links, limited physical security, bandwidth-constrained and energy-constrained operation.MANETS can be used for facilitating the collection of sensor data for data mining for a variety of applications such as air pollution monitoring and different types of architectures can be used for such applications. It should be noted that a key characteristic of such applications is that nearby sensor nodes monitoring an environmental feature typically register similar values. This kind of data redundancy due to the spatial correlation between sensor observations inspires the techniques for in-network data aggregation and mining. By measuring the spatial correlation between data sampled by different sensors, a wide class of specialized algorithms can be developed to develop more efficient spatial data mining algorithms as well as more efficient routing strategies. Also researchers have developed performance models for MANET by applying Queuing Theory.

II.RELATED WORKS

A Mobile Ad hoc Network (MANET) is made up from a set of MANET routers (MRs). These MRs organize and maintain a routing structure among themselves over dynamic wireless interfaces. As any Internet Protocol (IP) router, a MR may have an attached set of nodes. These nodes access the MANET via the MR to which they are attached.

Due, in part, to relative movements of MR and, in part, to environmental effects (especially wireless characteristics), the network topology and communication links in a MANET may change state more frequently than in fixed wired or fixed wireless networks. These attributes and others influence Internet Protocol (IP) design for MANETs. MANETs may in some cases operate as stand-alone networks, or they may be used to extend the wireless. Mobile range of a more fixed infrastructure network designs.

Routing has always been one of the key challenges in MANETs and the challenge becomes more difficult when the network size increases. Many routing protocols for MANETs have been proposed, which can be divided by flat routing and hierarchical routing. The protocol DSR and AODV are typically flat routing. However, when the network grows large and there are more mobile nodes in the network, it takes more size to keep the route path, as well as this path becomes so unstable that it will break at any time for flat routing. Some other routing algorithms for mobile ad hoc networks have been proposed to handle the mobile nodes. Position-based routing algorithms use some additional information about the physical position of the participating nodes through the use of the GPS or some other types of position service, which is not always available.

DSR over AODV Routing (DOA) is a lightweight hierarchical routing protocol which divides route into segments that we can take it as one dimension hierarchy. At a higher level, DSR is being used as a global intersegment routing protocol while AODV is being used as intra-segment routing at a lower level. It claims that once a route broken happens, it can be fixed locally at the level of a segment. Although its overhead is lower than strict hierarchical routing, the waypoint nodes in higher level are as unstable as other nodes, which cause an intersegment route repair happen frequently. So its performance in MANET is still restricted. In this paper, we use the same hierarchical model, while our main focus is on how to manage the cell (cluster) to make it more stable and maintain the overhead in a more acceptable level. Besides, we also manage to reduce the traffic through the cell-head to make it more sustainable.

III. MULTICLASS ROUTING

Power heterogeneity is common in mobile ad hoc networks (MANETs). With high power nodes, MANETs can improve network scalability, connectivity and broadcasting robustness. However, the throughput of power heterogeneous MANETs could be severely impacted by high-power nodes. To address this issue, we present LRPH, a *loose* virtual clustering Based *R*outing protocol for *P*ower *H*eterogeneous MANETs. In particular, to explore the advantages of high-power nodes, we develop a loose virtual clustering algorithm to construct a hierarchical network and eliminate unidirectional links. To reduce this interference raised by high-power nodes, we develop routing algorithms to avoid packet forwarding via high-power nodes. Via the combination of analytical modeling, simulations, and real-world experiments, we demonstrate the effectiveness of LRPH on improving the performance of power heterogeneous MANETs.

Mobility and Infrastructure

Infrastructure includes both physical assets such as highly specialized buildings and equipment, as well as non-physical assets such as the body of rules and regulations governing the various systems, the

financing of these systems, as well as the systems and organizations by which highly skilled and specialized professionals are trained, advance in their careers by acquiring experience, and are disciplined if required by professional associations (professional training, accreditation and discipline).

Mobility Model

In the performance evaluation of a protocol for an ad hoc network, the protocol should be tested under realistic conditions including, but not limited to, a sensible transmission range, limited buffer space for storage of messages, representative data traffic models, and realistic movement of mobile users. This page describes several mobility models that represent mobile nodes movements are independent of each other.
Communication and interference Models

In communications and electronics, especially in telecommunications, **interference** is anything which alters, modifies, or disrupts a signal as it travels along a channel between a source and a receiver. The term typically refers to the addition of unwanted signals to a useful signal.

CONCLUSION

In this paper advantage of different communication capabilities of heterogeneous nodes in many realistic mobile ad hoc networks presented multiclass routing models. An efficient media access control layer protocol-HMAC heterogeneous MANETs is also presented. Different of nodes have different transmission power, bandwidth processing capability, reliability and security. This reduces the number of routing hops and makes the routing more efficient and reliability. Large bandwidth, transmission ranges are more reliable. Extensive simulation tests demonstrate MC has very good performance, and performs, much better in terms of reliability, scalability, routing overhead as well as packet delay and throughput.

REFERENCES

[1] P. Gupta and P. R. Kumar, "The capacity of wireless networks," IEEE Trans. Inf. Theory, vol. 46, no. 2, pp. 388–404, 2000.

[2] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of ad hoc wireless networks," IEEE/ACM Trans. Networking, vol. 10, no. 4, pp. 477–486, 2002.

[3] M. Garetto, P. Giaccone, and E. Leonardi, "Capacity scaling in delay tolerant networks with heterogeneous mobile nodes," in ACM MobiHoc 2007, New York, USA, 2007, pp. 41–50.

[4] M. J. Neely and E. Modiano, "Capacity and delay tradeoff for ad hoc mobile networks," IEEE Trans. Inf. Theory, vol. 51, no. 6, pp. 1917– 1937, 2005.

[5] X. Lin, G. Sharma, R. R. Mazumdar, and N. B. Shroff, "Degenerate delay-capacity tradeoffs in ad-hoc networks with brownians mobility," IEEE/ACM Trans. Networking, vol. 14, no. SI, pp. 2777–2784, 2006.

[6] Y. Wang, X. Chu, X. Wang, and Y. Cheng, "Optimal multicast capacity and delay tradeoffs in manets: A global perspective," in IEEE Infocom 2011, Shanghai, China, 2011.

[7] P. Li, Y. Fang, and J. Li, "Throughput, delay, and mobility in wireless ad hoc networks," in IEEE Infocom 2010, San Diego, CA, USA, 2010.

[8] L. Ying, S. Yang, and R. Srikant, "Optimal delay-throughput trade-offs in mobile ad-hoc networks," IEEE Trans. Inf. Theory, vol. 54, no. 9, pp. 4119 – 4143, 2008. [9] S. Zhou and L. Ying, "On delay constrained multicast capacity of large-scale mobile ad-hoc networks," in IEEE Infocom 2010 mini-conference, San Diego, CA, USA, 2010.

[10] W. Huang and X. Wang, "Throughput and delay scaling of general cognitive networks," in IEEE Infocom 2011, Shanghai, China, 2011.

[11] H. Li, X. Liu, W. He, J. Li, and W. Dou, "End-to-end delay analysis in wireless network coding: A network calculus-based approach," in IEEE ICDCS 2011, Minneapolis, MN, USA, 2011.

[12] B. Liu, Z. Liu, and D. Towsley, "On the capacity of hybrid wireless networks," in IEEE Infocom 2003, vol. 2, San Francisco, USA, 2003, pp. 1543–1552.

[13] U. C. Kozat and L. Tassiulas, "Throughput capacity of random ad hoc networks with infrastructure support," in ACM MobiCom 2003, New York, USA, 2003, pp. 55–65.

[14] A. Agarwal and P. R. Kumar, "Capacity bounds for ad hoc and hybrid wireless networks," SIGCOMM Comput. Commun. Rev., vol. 34, no. 3, pp. 71–81, 2004.

[15] P. Li and Y. Fang, "Impacts of topology and traffic pattern on capacity of hybrid wireless networks," IEEE Trans. Mobile Computing, vol. 8, no. 12, pp. 1585 – 1595, 2009.

[16] P. Li, M. Pan, and Y. Fang, "The capacity of three-dimensional wireless ad hoc networks," in IEEE Infocom 2011, Shanghai, China, 2011.

[17] X.-Y. Li, "Multicast capacity of wireless ad hoc networks," IEEE/ACM Trans. Networking, vol. 17, no. 3, pp. 950–961, 2009.

[18] X. Mao, X.-Y. Li, and S. Tang, "Multicast capacity for hybrid wireless networks," in ACM MobiHoc 2008, Hong Kong, China, 2008, pp. 189–198.

[19] W. Huang, X. Wang, and Q. Zhang, "Capacity scaling in mobile wireless ad hoc network with infrastructure support," in IEEE ICDCS, 2010, Genoa, Italy, 2010, pp. 848–857.

[20] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting places from traces of locations," SIGMOBILE Mob. Comput. Commun. Rev., vol. 9, no. 3, pp. 58–68, 2005.

[21] Z. Yang and Y. Liu, "Understanding node localizability of wireless ad- hoc networks," in IEEE Infocom 2010, San Diego, CA, USA, 2010.

[22] G. Sharma, R. Mazumdar, and N. B. Shroff, "Delay and capacity trade- offs in mobile ad hoc networks: A global perspective," IEEE/ACM Trans. Networking, vol. 15, no. 5, pp. 981–992, 2007.

[23] M. Garetto, P. Giaccone, and E. Leonardi, "Capacity scaling in ad hoc networks with heterogeneous mobile nodes: The super-critical regime," IEEE/ACM Trans. Networking, vol. 17, no. 5, pp. 1522–1535, 2009.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Dynamic Storage Security in Cloud Computing

V. Sridevi [1], M.Geetha [2]

[1]Assistant Professor, Department of Computer Applications
Dr. N.G.P.Arts and Science College, Coimbatore-48

[2]Assistant Professor and Research Scholar, Department of Information Technology,
KG College of Arts and Science, Coimbatore.

ABSTRACT

It is not an easy task to securely maintain all essential data where it has the need in many applications for clients in cloud. To maintain our data in cloud, it may not be fully trustworthy because client doesn't have copy of all stored data. But any authors don't tell us data integrity through its user and CSP level by comparison before and after the data update in cloud. So we have to establish new proposed system for this using our data reading protocol algorithm to check the integrity of data before and after the data insertion in cloud. Here the security of data before and after is checked by client with the help of CSP using our ―effective automatic data reading protocol from user as well as cloud level into the cloud‖ with truthfulness. Also we have proposed the multi-server data comparison algorithm with the calculation of overall data in each update before its outsourced level for server restore access point for future data recovery from cloud data server. Our proposed scheme efficiently checks integrity in efficient manner so that data integrity as well as security can be maintained in all cases by considering drawbacks of existing methods

Index Terms— Data management in cloud, Data integrity, TPA, Cloud multiple server, Dynamic operation. This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an IEEE conference. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.

Keywords— Data management in cloud, Data integrity, TPA, Cloud multiple server, Dynamic operation.

I. INTRODUCTION

Cloud computing is a virtualized resource where we want to store all our data with security measurement so that some application and software can get full benefits using this technology without any local hard disk and server for our data storage. These services are broadly divided into three categories as

1) Infrastructure-as-a-Service. 2) Platform-as-a-Service and 3) Software-as-a-Service [1] [7]. It is not so efficient by using a few technologies such as flexible distributed scheme, Reed Solomon technique, and BLS algorithm [8] to give more integrity to the operation like change, removal, and add of data from cloud server. So to offer yet sophisticated and efficient integrity to cloud data so that all applications, software and highly valuable things can‘t be affected and modified according to unknown person‘s challenge from un-trusted cloud we have to provide more integrity method apart from previous methods with the assist of CSP who maintains our data from cloud servers from their IP address domain [3]. But since we may not have copy of outsourced data CSP can behave falsely regarding the status of our outsourced data. And since we don‘t have any physical data control we have to fully depend on CSP for cloud data access without any separate login for our self access in data management. So we have some limitation in this view so that CSP only can take more rights to manage our data with his IP address domain. But here we need to keep up data integrity for our cloud data in efficient manner for our own usage whenever there is a want for that and must have some integrity measurement for our data storage to protect data from internal and external attack including byzantine as well as malicious attack. And in this case, there are some different dealings such as file distribution algorithm, error location correctness and data authentication [8] using distributed protocol for its integrity as far as existing system is concerned. But all these appear from existing system with a few restrictions without learning data integrity in competent manner using

effective automatic data reading protocol from user as well as cloud level before and after the data append into the cloud. And also we have done multi-server data comparison algorithm for every data upload for making efficient way for data integrity. If there is any server failure, then the data can be recovered repeatedly in cloud server using this scheme.

II. ACTIVITIES OF THE CLOUD

A. Devoted Property Assertion

All resources have high-quality assurance in its multi cloud servers with the help of strong platform. The mentioned possessions of RAM, CPU, and Bandwidth capacity [4][5] in network produce sufficient atmospheres for its concerned usage. Thus the activity of could server can‘t ever opposite to the current world while considering its level of integrity and security even though there is some little bit of conflict that we want yet to recover from its already existing level.

B. Excess Data Storage

All cloud server storage resources are managed by high-performance and high-availability storage area network. Many cloud solutions operate on local disks from the host system, which means any computing or storage failure can result in down time and potential data loss. As cloud servers are autonomous, if there happens any server crack in stored data, these can be protected against internal and external attacks.

C. Client Support and Stability

It clearly tells us that user can do the increase and decrease of the data capacity in the cloud server with the help of CSP (cloud service provider) in his request. This storage level must be with flexible and durability condition as far as its entire design or structure is concerned. Thus it should be claimed extra storage space concerning future process in data exchange.

D. Pay Costs for Access

Customers have to be charged according to their usage and their data storage in cloud storage apart from allocated space. Customers can charge depending on the usage of network access for data exchange with good bandwidth [4] [5] as well as data storage. For example, it is just like when utility company sells power to consumers and telephonic company offering utilities for their services.

E. Efficient Dynamic Computing Environment

Data should be in a self-motivated computing infrastructure. The main concept in this dynamic environment is that all standardized and scalable infrastructure should have dynamic operation such as modification, append, and delete. The cloud platform which has virtualized conditions also should have some specific independent environment.

F. Accurate Choice of CSP

To get excellent service from multiple servers, good service providers are important to be considered and selected. So much care must be taken in this respect so that the CSP itself can be elastic with clients in order to get accessed with all places (anywhere and anytime). It has the following benefits such as, 1) Cost savings- To save expenditure among IT capabilities. 2) Trustworthiness- Data back up by CSP in cloud servers if system is stolen and loses the data itself. 3) Scalability on requirement- Whenever there is a need for data accessing, user can be easily accessible to that anytime and anywhere. 4) Protection expertise- Cloud service providers in general have more skill securing networks than individuals and IT personnel whose networks are usually associated to the public Internet anyway. 5) All over Access- It allows users ubiquitous access to computing power, storage, and applications.

III. PROBLEM REPORT AND SYSTEM STRUCTURE

This has following system formation for its structure as 1) User 2) Multiple cloud servers 3) Cloud service provider 4) Data reading protocol 5) Server access point.

A. A. Making Server Access Point and Time Intimation

To keep the data away from server failure in every data conclusion by unauthorized person or any internal and external attack coming within CSP address domain, one access point or restore point in every update is given to the cloud server when client does some delete, modification, and append in his will. It is done in the time of data comparison in every update for the data by user with the help of CSP. This is a new technique we have introduced in our system design. The reason for which we want to do this restore point is that if there is any server crash then this restore access point helps a lot to recover the whole thing that we have lost. This process is proceeded by fixing one access point to cloud server database in update or data upload (it is done with the help of CSP, because we don‘t have any replica of our outsourced data) when we finally terminate our data exchange in cloud server. Also since we don‘t have any physical custody of our data in cloud server we can‘t have any separate login access using cryptographic key. It is a major disadvantage to our cloud server to maintain our individual or group data. In coming days it can be improved with the help of CSP. When user uploads the data into the cloud automatically these data are included and created for access point and here one specific restore point is done in its cloud server for effective recovery purpose.

B. Optional Third Party Auditor

To reduce the work load of the user, (see Figure 1) TPA [1] [8] has the delegation for the data checking with a few limitation so that he can‘t modify content of the data in his auditing. And TPA pays due care on storage correctness verification.TPA auditing is in the manner of privacy preserving concept so that verification can be done in separate manner alone from any interaction by others. In the existing system, distributing protocol is used for this purpose. Here TPA does not have special authority for data integrity checking. However we don‘t give more importance to TPA to such a great degree but a small bit only since our proposed system pays more concentration on client‘s access for full security.

C. What the Existing System Has?

Many expertise‘s express us that whenever opponent tries for any attempt to append, delete and modification for cloud data from its storage level it is protected from that (byzantine problem, malicious, data modification attack, and also server failure) [8] using flexible distributed storage integrity auditing mechanism with distributed protocol. And it tells us that whatever changes has occurred by the above stated failure it is cleared only using specified file distribution and error correctness method. Typically it concentrates mostly on TPA auditing other than the client. Token format is followed for data arrangement in array format for their task to be performed. All Security measurement is processed through state-of-the-art.

D. Disadvantages of Existing System

Even though it concentrates more on time consuming for user almost user has to depend on TPA for data integrity whenever he doesn‘t have time for auditing and he automatically can‘t have the any stored data size after any such malicious, byzantine and system failures to know about data modification, delete and append in his own knowledge. But it does not tell about its (data) fixed storage capacity for user‘s stored data in cloud before and after in cloud storage area or in server (how much of data has been stored in server for particular clients in his own allotted storage server area). So it is a major issue to user and also almost users have to depend on CSP for extensive security analysis and depend partially on TPA. It doesn‘t tell effective way for server failure. It doesn‘t make efficient rout for data integrity whenever user want required data from his earliest storage in cloud server. But here complete access is going to CSP. So CSP can behave in its own way by hiding any loss of data since existing system doesn‘t tell about the weight or value of stored data using any algorithm.

E. Our Proposed Model

We suppose that CSP allotted space is a major concern for our data maintenance in all manner for dynamic operations. All outsourced data and data entering into the cloud is measured using reading protocol algorithm. So in order to keep integrity of overall data, we have to use ―data reading protocol from user as well as cloud storage level before and after the data adding into the cloud server area‘ and another ―multi-server data comparison algorithm for every data upload for the purpose of data recovery management‖ for our proposed achievement. When server failure occurs in cloud entire data may be affected so that user can‘t foresee data‘s trustworthiness in its whole atmosphere depending on variety of situation or CSP‘s process to hide the loss of data. Taking into account these, we have modeled and proposed automatic protocol and server data management algorithm to know about entire data exchange before and after the data insertion into multiple cloud server. And user can know if there has been done any change, remove, and attach operations that can occur for data from its storage level with the help of our proposed scheme that never has been told. It can be effectively administered by clients from appropriate efficient data reading protocol from cloud server position.

F. Advantages of Proposed Model

By processing the integrity of data using data reading protocol and data management algorithm after and before the entering of data into the cloud, user can assure that all data in cloud must be in protected condition for its trustworthiness. So easily the actual size of stored data before and after in cloud is maintained even though the user himself has done any modification, deletion, and update for his own purpose by using proposed scheme. These processes are carefully done using our proposed scheme. So here user takes full control and process on the data stored in cloud apart from TPA and we can give strong assurance and protection to the data stored in multiple cloud server environment. To avoid server failure and any unexpected error we should put one server restore point in cloud server database for efficient data back up or restore using multi server data comparison method. It is major advantage of our proposed system. This process is done with the help of CSP for cloud database process since we have physical data possession in cloud server.

V.OUR PROPOSED DATA BLOCK DYNAMIC PROCESS

To ensure the assurance of the data we can do the operations such as append, deletion, and update. These are the dynamic data operations to be done in the cloud area by user.

A. Append Operation in Server Block

We may assume that there is any size of GB space allotted by CSP user‘s requirement for any application

purpose. Then first, this size is calculated and compared using our technique. It is clearly mentioned in our algorithm specification. In the comparison the storage cloud area is confirmed that it does not have any data in its position for strong integrity. Also the operations such as add, change, and removals by client are processed by space measurement scheme for effective identification of data integrity in cloud database. So if there is any such modification by attack then client can give assurance to the data integrity by successfully following. It is very efficient method in our proposed design compared to any other such type.

B. Deletion Operation in Server Block

First, we want to compare value from existing cloud server. Then, this deletion operation depends on user‘s attempt on his data stored in cloud server using his login operation. The following operation is performed in this deletion operation. If there is number of servers for data selection while deleting the data stored the particular storage server is considered for this. The required data can be deleted using above mentioned algorithm.

C. Update Operation in Server Block

This operation is finally finished after completion of needed action by user. And for update, the above derived algorithms are taken into the account. For new data update, each data block should be updated automatically from already existing value to new updated value. So if we consider this data block as one array formation then, the result is as S= (Si ± Di) depending on the user‘s operation on data. It may be positive or negative depending on the update operation to be performed.

VI. CONCLUSION

From our study we have good conclusion of the integrity and security of cloud level data when user do some update such as attach, removal, and change at own desire. So here user mostly has the situation to believe the service provider in many manners. Understanding all these situations, we have handled new proposed technique for cloud storage and. This is the modern way we have used in our paper different algorithm compared to previous related paper in this respect for cloud management. From this protection for cloud data, user can be strong belief for his uploaded data for any future purpose or his any other related process without worry. Here complicated, internal, external and malevolent attack is known by our proposed scheme in efficient manner by storage data measurement (i.e., since there can be some modifications in cloud data). Thus our main idea is to give integrity to the cloud storage area with strong trustworthiness so that user can feel free of worry for his uploaded data in his allocated space. *Here* our scheme ensures for any extra inclusion of unwanted bits or related things in cloud area so that they can be so easily found out by our data measurement concepts in efficient manner. And it finds out how much of changes have occurred there in its cloud

VII. FUTURE ENHANCEMENT

Here we leave more ways as Future enhancement to process for maintaining security and integrity of data using read and write protocol for data calculation from cloud data storage in days to come so that user can identify inserting the attempt of different data having same weight in un-trusted cloud server. This read and protect protocol is efficient to identify any data modification into the server with accurate reading and protecting capacity automatically when such attempt is made by known one. In our future study we also have planned to 6

implement the locking protocol in cloud data storage with the help of CSP for data update. It can give clear security to user‘s own data when users complete his requirement.

REFERENCES

[1] Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing Qian Wang1, Cong Wang1, Jin Li1, Kui Ren1, and Wenjing Lou2-Springer-Verlag Berlin Heidelberg -2009.

[2] Cloud security defence to protect cloud computing against HTTP-DoS and XML-DoS attacks, Ashley Chonka,YangXiang n, WanleiZhou, AlessioBonti, Elsevier.

[3] A Multiple-Replica Remote Data Possession Checking Protocol with Public Verifiability-2010 Second International Symposium on Data, Privacy, and E-Commerce.E. H. Miller, ―A note on reflector arrays (Periodical style—Accepted for publication),‖ *IEEE Trans. Antennas Propagat.*, to be published.

[4] Amazon.com, ―Amazon s3 availability event: July 20, 2008,‖Online at http://status.aws.amazon.com/s3-20080720.html, July 2008.

[5] Sun Microsystems, Inc., ―Building customer trust in cloud computing with transparent security,‖ Online at https://www.sun.com/offers/details/suntranspar.xml, November 2009. M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.

[6] Addressing cloud computing security issues, Future generation computer systems(2011) www.elsevier.com/locate/fgcs.

[7] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, ―Provable data possession at untrusted stores,‖ in Proc. of CCS‘07, Alexandria, VA, October 2007, pp. 598–609.

[8] Towards Secure and Dependable Storage Services in Cloud Computing Cong Wang, Student Member, IEEE, Qian Wang, Student Member, IEEE, Kui Ren, Member, IEEE, Ning Cao, Student Member, IEEE, and Wenjing Lou, Senior Member, IEEE- 2011.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Survey of Spectrum Sensing issues and attacks in Cognitive Radio Networks

K.Gomathy [1],P.Mohanasundhari[2], R.Rajesh kanna [3]

[1 2 ,3]Assistant Professor, Dr.N.G.P Arts and Science College, Coimbatore

[2] Assistant Professor, Kongu Arts and Science College,Erode

ABSTRACT

Cognitive Radio (CR) is a novel technology that promises to solve the spectrum deficiency problem by allowing secondary users to coexist with primary users without causing interference to their communication. Although the operational aspects of CR are being explored vigorously, its security aspects have gained little attention. In this paper, a brief overview of the CR technology is provided followed by a detailed analysis of the security attacks targeting Cognitive Radio Networks (CRNs) along with the corresponding mitigation techniques. We categorize the attacks with respect to the layer they target starting from the physical layer and moving up to the transport layer. An evaluation of the suggested countermeasures is presented along with other solutions and augmentations to achieve a secure and trusted CRN. The spectrum sensing problem has gained new aspects with cognitive radio and opportunistic spectrum access concepts. It is one of the most challenging issues in cognitive radio systems. In this paper, a survey of spectrum sensing methodologies and security attacks in cognitive radio is presented.

Keywords Spectrum sensing, Attacks, CRN

Introduction

The ever increasing demand of spectrum due to the rapid introduction of novel wireless applications has led the Federal Communication Commission (FCC) to approve in September 2010 new rules to allow unlicensed users to utilize the spectrum reserved for wireless broadband services (300MHz and 400MHz). The technology developed to take advantage of this unused spectrum is Cognitive Radio Networks (CRNs) which are intelligent networks that adapt to changes in their environments to make a better use of the radio spectrum. CRNs help solve the problem of spectrum shortage by allowing unlicensed users to use primary systems without interference. This technology allows the coexistence and sharing of licensed spectrum resources between two types of users, licensed and unlicensed.

The successful deployment of CRNs includes the correct construction and maintenance of security measures to combat attacks launched against them. We categorize the attacks on CRNs into four major classes: Physical Layer attacks, Link Layer attacks (also known as MAC attacks), Network layer attacks, and Transport Layer attacks. In Physical Layer, we discuss Primary User Emulation (PUE),

Objective Function Attack, and Jamming. In Link Layer, we discuss Spectrum Sensing Data Falsification (SSDF), Control Channel Saturation DoS Attack (CCSD), and Selfish Channel Negotiation (SCN). In Network Layer, we mainly discuss the routing attacks that are relevant to CRNs, for instance, HELLO Flood attack and Sinkhole attack. In transport Layer, we discuss the Lion Attack. Some of these attacks might target different layers such as jamming which can be done in either the physical or MAC Layers. After presenting each attack we discuss in details the techniques used to mitigate it. We then evaluate these countermeasures showing their strengths and weaknesses.

Being the focus of this paper, spectrum sensing by far is the most important component for the establishment of cognitive radio. Spectrum sensing is the task of obtaining awareness about the spectrum usage and existence of primary users in a geographical area. This awareness can be obtained by using geolocation and database, by using beacons, or by local spectrum sensing at cognitive radios [4]–[6]. When beacons are used, the transmitted nformation can be occupancy of a spectrum as well as other advanced features such as channel quality. In this paper, we focus on spectrum sensing performed by cognitive radios because of its broader application areas and lower infrastructure requirement. Other sensing methods are referred when needed as well. Although spectrum sensing is traditionally understood as measuring the spectral content, or measuring the radio frequency energy over the spectrum; when cognitive radio is considered, it is a more general term that involves obtaining the spectrum usage characteristics across multiple dimensions such as time, space, frequency, and code. It also involves determining what types of signals are occupying the spectrum including the modulation, waveform, bandwidth, carrier frequency, etc.. However, this requires more powerful signal analysis techniques with additional computational complexity.

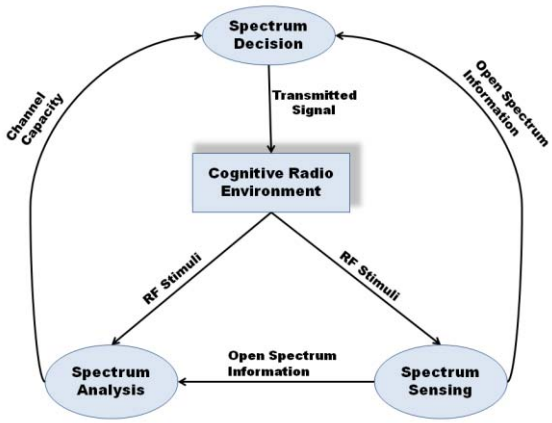


Figure 1: Cognition cycle

an be considered a special kind of Ad Hoc network, most of the attacks targeting Ad Hoc networks can also target CRNs. In this survey, we analyze the attacks that are most relevant to CRNs.

It is important to note that there already exist some surveys on CRNs [12-13], but they have many weaknesses in the sense that they miss to address some very important attacks, they are outdated, and most importantly none presents an evaluation study of the various countermeasures.

Any solution suggested to counter CRN attacks should abide by the FCC requirement which states that “no modification to the incumbent system should be required to accommodate opportunistic use of the spectrum by secondary users” [14]. Having this requirement in mind, any security solution suggested to protect or thwart an attack on CRN should be introduced to the secondary user system, not the primary one.

Defending against HELLO Flood Attacks

To countermeasure the HELLO flood attacks, a symmetric key should be shared with a trusted base station [42]. The base station will act as a Trusted Third Party as in Kerberos and facilitate the establishment of session keys between parties in the network; in order to protect their communication. Consequently, two nodes may use the session key to verify each other's identity; as well as, authenticate and encrypt the link between them. Now, to prevent an attacker from creating a session key with every node on the network, the number of shared keys must be limited. In addition, a node claiming to be the neighbor of so many nodes in the network should raise an alarm. Symmetric key algorithms are suggested because they are known to be faster and have lower overhead on system resources.

General Techniques to Defend against Network Layer Attacks

In general, one can defend against routing attacks by using a secure routing protocol, such as Secure Efficient Ad hoc Distance vector routing protocol (SEAD) [43]. SEAD protects against denial of service attacks as it uses a one way hash function instead of asymmetric encryption to prevent attackers attempts to cause other nodes to use more network bandwidth or processing time. The protocol operates as the vector routing protocol, and the design is based on Destination-Sequences Distance-Vector protocol (DSDV).

Another effective mechanism to defend against routing attacks is to use a cross layer solution to make the transmission more efficient [44]. It suggests that, instead of router's direct decision, the routing algorithm and spectrum management should be considered together to make decisions for the channel scheduling.

Transport Layer Attacks

As with the other layers, the Transport layer in a CR node is also vulnerable to many of the attacks that target wireless Ad Hoc networks in general, for instance, the JellyFish attack [45]. In what follows, we only consider a transport layer attack named Lion Attack [46] because of its close relevance to CRNs.

Lion Attack

The Lion attack uses the primary user emulation (PUE) attack to disrupt the Transmission Control Protocol (TCP) connection. The Lion attack can be considered a cross-layer attack performed at the physical link layer and targeted at the transport layer where emulating a licensed transmission will force a CRN to perform frequency handoffs and thus degrading TCP performance. When a PUE attack is performed, all SUs have to do frequency handoff in order to free the channel for the primary user. When this handoff takes place, TCP will not be aware of the handoff and will keep creating logical connections and sending packets without receiving acknowledgments. The TCP segments will then start to timeout and consequently TCP retransmits them with an increased timeout value. As a result, the retransmission timer backs off doubling the value, resulting in delays and packet loss. Additionally if an attacker can intercept the messages, it can predict the frequency band tested in a handoff, and claim it using PUE resulting in a total network starvation.

Defending against Lion Attack

To Mitigate the Lion attack, Hernandez-Serrano et al. suggest a mechanism that starts by making the TCP protocol aware of what is happening in the physical layer by employing cross-layer data sharing between physical/link and transport layers. This way, the CRN devices will be able to freeze TCP connection parameters during frequency handoffs and adapt them to the new network

conditions after the handoff. To secure the control data in order to prevent the attacker from eavesdropping current and future actions of the CRN, a group key management (GKM) can then be used to allow CRN members to encrypt, decrypt and authenticate themselves. Finally, a cross-layer IDSs specifically adapted to CRNs can be used as a technique to find the attack source if it still exists.

Finally, cognitive radios must have some common sense [1]. Policies must be defined to cover all scenarios. In addition, some sort of cooperation between the different cognitive radios can be beneficial. In [1] a technique called particle swarm optimization (PSO) is mentioned. Each cognitive radio is a particle, and each has its own idea about what is the best behavior in a particular situation. However, this behavior is not dependent solely on its own idea, but a weighted average of all the ideas in the network.

SECURITY REQUIREMENTS

Cognitive radios are subject to many of the same types of attacks as other wireless networks. In addition, the cognitive radio is subject to attacks specific to the radio itself. These attacks may target the spectrum sensing function by changing the spectrum environment, the decision making function by manipulation of parameters of the objective function, or the learning engine by providing false data about the environment
that the learning radio will use in the future to make incorrect or inefficient decisions.

The core principles of secure communications are based on the CIA triad:

1) Confidentiality refers to protecting information such that there is no unauthorized disclosure to individuals or systems. Confidentiality of data is required to maintain the privacy of the data owner. The data owner can include a bank storing credit and balance information about a customer, as well as the customer.

2) Integrity refers to maintaining and assuring the accuracy and consistency of stored data over its life cycle. Integrity means that data cannot receive unauthorized or undetected modification. In addition, integrity is violated when a message is modified while in transit.

3) Availability refers to the ability of the authorized user to obtain information or transfer messages as needed. Availability applies to the system that stores the data, the security controls that protect the data, and the communication used to retrieve the data. In addition to the triad, two other principles must be applied.

1) Authenticity refers to validating the identities of the parties involved in a communication.

2) Non-repudiation refers to the inability of a party to deny sending or receiving a transaction.

By classifying threats we can better determine threat severity, precautionary methods, and recovery strategies. Additionally, understanding the similarities between the threats can help us apply knowledge about previous attacks on other technologies to cognitive radio networks. The following framework provides a classification system for all cognitive radio network threats. The threats are classified according to the protocol layer upon which the attack is performed: physical layer, data link (or MAC) layer, network layer, application layer, and cross layer.

Cross layer attacks are those in which the attack is launched utilizing one layer while the attack targets another layer. Basing the classifications upon protocol layers utilizes terminology already used in wireless communication security while simultaneously describing for the reader the attack

vector. We start our discussion at the bottom of the layer stack and move upwards.

PHYSICAL LAYER

The physical layer is the lowest layer of the protocol stack, providing an interface to the transmission medium. The physical layer consists of anything that is used to make two network devices communicate, such as the network cards, fiber, or, as in the cognitive radio network framework, the atmosphere. The operation of the cognitive radio network is more complicated than other wireless communication networks because the cognitive radio uses the frequency spectrum dynamically. Following are network attacks aimed at disrupting communication by targeting the physical layer of the cognitive radio network.

Data Link Layer

The data link layer in the context of 802.11 consists of two sublayers. These are the Logical Link Control (LLC) and Media Access Control (MAC). The Logical Link Control layer manages traffic flow and error control over the physical medium. The Media Access Control layer is designed to support multiple users on a shared medium within the same network. Each computer is assigned its own unique MAC address. The attacks described below affect the MAC sublayer.

Network Layer

The network layer provides the ability to route data packets from a source node on one network to a destination node on another network, while maintaining quality of service. It also performs fragmentation and reassembly of packets, if required. The cognitive radio network shares security issues with the classic wireless communication networks due to the three shared architectures of mesh, ad hoc, and infrastructure. Cognitive radio networks also share similarities with wireless sensor networks. These include multi-hop routing protocols and power constraints. In addition, there are special challenges faced by cognitive radio networks due to the required transparency in the existence of the network activities to the primary user. Routing in the cognitive radio network is further complicated by the requirement of the radio to vacate the frequency when the primary user is sensed as present. Cognitive radio security vulnerabilities are therefore also inherited from these architectures.

Application Layer

The application layer is the layer closest to the end user. The user and the application layer interact with the application software. The application layer is responsible for determining the resources available, synchronizing communication, and identifying the communicating devices. Cognitive radios require a greater processing power and memory capacity than the traditional smart phone. This is because of the extra tasks performed by the cognitive radio, such as spectrum sensing and learning.

Transport Layer

The transport layer responsibilities include flow control, congestion control, and end-to-end error recovery. The transport layer in the cognitive radio network is subject to many of the vulnerabilities that plague wireless ad hoc networks.

SECURITY THREATS & VULNERABILITIES

Rogue Base Station Attack:

The attackers lie in the center of the Base Station Offer (BSO) and Base Station Renter (BSR). The attacker could take off himself as an occupant and can send back a resource return message to the offer. In the same way the attacker could pose as offer and request for resource allocation. Attacker can generate the BS-IDs and can forge the network.

Attacker can get BS-IDs and negotiated channels between the offer and renter at the time of resource sharing. The attacker abuses this information and sends bogus messages.

Replay Attack: The attacker can capture the packet and resend these packets maliciously after passing some delay for mismanagement of the network characteristic & resources. Attacker can retransmit the packet after certain time to cause DoS.

Selfish Attack: The attacker can access the spectrum with high priority. He occupies the spectrum resource as he wants.

Malicious Attack: The opponent slows down the unlicensed user from spectrum usage and caused denial of services (DoS).

Cheat Attack: The attacker increase his effectiveness function as well as at the same time decrease contestant's profit.

Misbehaving: CR didn't follow any general rule for sensing and managing of range of the spectrum.

Denial of Service (DoS): During the period of sensing any attacker / hacker can flood the spectrum with inconsistent data to show off that the range is unavailable for the same because of this the sensing during is the most vulnerable to DoS. The major objective of DoS attack is to put the burden on the resources and to stop the utilization of the resources for the nodes which are also present in the network.

Incumbent Emulation (IE) Attacks : Under the Incumbent Emulation the user tried to get the priority on the other secondaries by sending signals which could help to emulate the features of an incumbent. The impact of Incumbent Emulation attack is the genuine secondary's abilities for differentiate the signals of attacker with the actual incumbent signals during the sensing period.

Spectrum Sensing Data Falsification (SSDF) Attacks: The hacker can spoof or mask primary user and transmit incorrect spectrum deduct sensing outcomes to a data collector, which is a major source of incorrect spectrum detection decision taken by the data collector

Conclusion

In this paper, we described the most recent and important attacks targeting CRNs. We classified them

according to the layer they operate on and presented their existing countermeasures. We then evaluated all the countermeasures giving each one a grade that presents its effectiveness. According to these evaluations, we suggested to combine the countermeasures that we think will produce the ultimate secure CRN. Such a suggestion should be normally supported by simulation results, but we keep this as part of our future work. We also overviewed the works that suggest building from scratch security frameworks for CRN.

The attacks were presented according to the layer of the protocol stack under assault. The special category of Cross Layer attacks was appended to the layer list to incorporate attacks launched at one layer, with the impact targeted at another layer. The layered presentation approach allowed us

to utilize vocabulary and ideas already associated with the protocol stack.

References

[1] T. Charles Clancy and Nathan Goergen, Security in Cognitive Radio Networks: Threats and Mitigation, International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom), Singapore, May, 2008, pp.1-8.

[2] Kwang Cheng Chen, Y. J. Peng, Neeli Rashmi Prasad, Y. C. Liang and Sumei Sun, Cognitive Radio Network Architecture: part I -- General Structure, Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, Suwon, South Korea, January, 2008, pp.114-119.

[3] Vinod Sharma and ArunKumar Jayaprakasam, An Efficient Algorithm for Cooperative Spectrum Sensing in Cognitive Radio Networks, Proceedings of National Communications Conference (NCC), Guwahati, India, January, 2009.

[4] Cognitive Radio Ad Hoc Networks, Broadband Wireless Networking Lab, School of Electrical and Computer Engineering, Georgia Inst of Tech. URL: http://www.ece.gatech.edu/research/labs/bwn/CRAHN/projectdescription.html

[5] Wenjing Yue and Baoyu Zheng, A Two-Stage Spectrum Sensing Technique in Cognitive Radio Systems Based on Combining Energy Detection and One-Order Cyclo-Stationary Feature Detection, Proceedings of the 2009 International Symposium on Web Information Systems and Applications (WISA'09), Nanchang, China, May, 2009, pp.327-330.

[6] Rajesh K. Sharma and Jon W. Wallace, Improved Spectrum Sensing by Utilizing Signal Autocorrelation, Proceedings of IEEE Vehicular Technology Conference, Barcelona, Spain, April, 2009, pp.1-5.

[7] Ruiliang Chen, Jung-Min Park and Jeffrey H. Reed, Defense against Primary User Emulation Attacks in Cognitive Radio Networks, IEEE Journal on Selected Areas in Communications, Vol.26, No.1, 2008, pp.25-37.

[8] Huahui Wang, Leonard Lightfoot and Tongtong Li, On PHY-Layer Security of Cognitive Radio: Collaborative Sensing under Malicious Attacks, 44th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, March, 2010, pp.1-6.

[9] Eric Wong and Rene Cruz, On Physical Carrier Sensing for Cognitive Radio Networks, Forty-Fifth Annual Allerton Conference on Communication, Control, and Computing, Allerton House, UIUC, IL, September, 2007.

[10] Bertrand Mercier, Viktoria Fodor, Ragnar Tobaben et al., Sensor Networks for Cognitive Radio: Theory and System Design, ICT Mobile Summit, Stockholm, Sweden, June, 2008.

[11] Tevfik Yucek and Huseyin Arslan, A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications, IEEE Communications Surveys & Tutorials, Vol.11, No.1, 2009, pp.116-130.

[12] Olga León, Juan Hernández-Serrano and Miguel Soriano, Securing Cognitive Radio Networks, International Journal of Communication Systems, Vol.23, No.5, 2010, pp.633-652.

[13] Xueying Zhang and Cheng Li, The Security in Cognitive Radio Networks: A Survey, Proceedings of the 2009 ACM International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly (IWCMC '09), New York, 2009, pp.309-313.

[14] Yiyang Pei, Ying-Chang Liang, Lan Zhang, Kah Chan Teh and Kwok Hung Li, Secure Communication Over MISO Cognitive Radio

Channels, IEEE Transactions on Wireless Communications, Vol.9, No.4, 2010, pp.1494-1



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

SPILT MULTIPATH ROUTING WITH MAXIMALLY DISJOINT PATHS IN AD-HOC NETWORKS

Gomathy.K[1], Hemapriya.K.E[2]

[1]Assistant Professor Department of Computer Applications

[2] Assistant Professor Department of Management Studies

Dr.N.G.P. Arts and Science College- Coimbatore-48

ABSTRACT

Routing has been the most focused area in ad hoc networks research. On-demand routing in particular, is widely developed in bandwidth constrained mobile wireless ad hoc networks because of its effectiveness and efficiency. Most proposed on-demand routing protocols however, build and rely on single route for each data session. Whenever there is a link disconnection on the active route, the routing protocol must perform a route recovery process. In QoS routing for wired networks, multiple path routing is popularly used. We propose an on-demand routing scheme called Split Multipath Routing (SMR) that establishes and utilizes multiple routes of maximally disjoint paths. Providing multiple routes helps minimizing route recovery process and control message overhead. Our protocol uses a per-packet allocation scheme to distribute data packets into multiple paths of active sessions. This traffic distribution efficiently utilizes available network resources and prevents nodes of the route from being congested in heavily loaded traffic situations. We evaluate the performance of our scheme using extensive simulation.

INTRODUCTION

An ad hoc network is an autonomous system of multi hop, wireless nodes that does not require base stations any fixed infrastructure. It is characterized by dynamic topologies,bandwidth-constrained, variable capacity links, energy constrained operation and limited physical security. The lack of infrastructure, in combination with multi hop connections and constantly changing topology pose difficult challenges on the routing protocol; foremost among them is how to deliver data packets while incurring the least routing overhead possible. SMR uses cooperative packet caching and shortest multi path routing to reduce Packet loss due to frequent route breakdowns. At present, caching is used not only in single-processor environments but also in distributed systems. Although additional storage overhead is required for the data cache and multiple routes, this technique, which we call Split Multipath Routing, can reduce packet loss due to route breakdowns. For cooperative packet caching to be effective, every node must maintain at least two routes to every active destination. One major problem of multiple path routing is the large routing overhead generated during route search and maintenance.In this paper, we introduced data packet caching in the context of the ad hoc networks. Performance improvements in terms of higher packet delivery, lower delays and reduced routing overhead are obtained due to the temporal locality of dropped packets. Furthermore, there is a high probability that the alternative route is always stale, as it is never used until the primary route fails. SMR cleverly avoids this problem by using all the available routes in a round-robin fashion. A multiple path route discovery mechanism suitable for ad hoc networks is also developed. The discovery mechanism selects non-disjoint shortest multi path routes of equal distance. We have shown that having at most two routes for every destination

is already the optimum. Having more routes does not give any additional benefit. Obviously, a five packet data cache and two routes per destination configuration do not entail a large overhead.

One major problem of multiple path routing is the large routing overhead generated during route search and maintenance. A multiple path protocol generates more than 50 times the routing overhead of AODV and DSR.

We now consider this situation: Suppose that some node I is along some route from source h to destination j. Suppose further that whenever a node fails to forward a data packet to its next hop, it drops the packet and sends a route error message to the source using the reverse route. Observe that whenever i receive a route error from a downstream node k, the error indicates the dropping of a data packet that source h has recently sent and i has recently forwarded. Clearly, dropped packets exhibit the property of temporal locality, that is, ‘‘a dropped packet is a recently sent packet’’. Therefore, if i has a buffer for caching forwarded data packets, even if the buffer is small, there is a big probability that the packet is still in that buffer. And if i have some other route to j, it can ‘‘salvage’’ the dropped packet and need not forward the route error upstream.

CHARACTERISTICS OF COOPERATIVE PACKET CACHING

Two kinds of locality have been observed - temporal and spatial. Spatial locality is the property whereby an access to a memory location indicates that a nearby location will very likely to be accessed in the near location indicates that the same location will very likely to be accessed again soon. It must be emphasized that without these properties,Every node maintains two data structures: a route cache to contain forwarding information; and a route request cache for storing recently received and processed route requests.Cooperative packet caching enables more nodes to salvage a dropped packet, or in essence, packet salvaging is distributed.

Over the last few years, many routing protocols for ad hoc networks have been proposed. In reactive routing, a node declares a link to a neighbor as down if it fails to forward a packet to this neighbor. This may cause one or more packets (at the node that encountered the forwarding failure) to become undeliverable. To avoid dropping such packets, DSR proposed an optimization known as ‘‘packet salvaging’’. In this optimization, the node that encountered the forwarding failure searches its route cache for alternative routes. If a route is found, undeliverable packets are forwarded through the alternative path. However, when no alternative path is found, these undeliverable packets are simply discarded. It is evident that in situations where link failures are frequent, this simple mechanism may result in increased packet loss, as there is no guarantee that packet salvaging will always be successful. In existing reactive routing protocols, only the node encountering the error can salvage or retransmit a data packet.

PROPOSED WORK

In our paper, we introduce Split Multipath Routing, a technique that exploits temporal locality in dropped packets, aimed at reducing packet loss due to route breakage. Every node maintains a small buffer for caching data packets that pass through it. When a downstream node encounters a forwarding error, an upstream node with the pertinent data in its buffer and alternative route can retransmit the data. For this strategy to be effective, nodes must store multiple routes to every active destination. Hence, we propose a simple route discovery mechanism that selects the shortest multipath routes.

ROUTE DISCOVERY

SMR operates on-demand, that is, node i maintains Sij only if there data packets to be sent to j. Its route discovery is similar to the ‘‘diffusing computations’’: given a direct acyclic graph (DAG), each node computes its distance based on the distance reported by the downstream nodes and reports its distance to its upstream nodes.

A source node h initiates route discovery when it has data to send to j but it has no available route. Node h floods the network with a RREQ (route request message) for j. This establishes a DAG rooted at h. When j receives a RREQ, it sends back a RREP (route reply message) to i through some nodes that is a subset of the DAG rooted at h.

Every RREQ from h to j has a forward count fc field, which is initialized to zero by the source and incremented by one every time the message is retransmitted. The first time i receives a RREQ from h to j, it initializes minfchji to fc and Phji to the previous hop of the message. Every time i receives a request with fc = minfchji (meaning the request traversed a path of the same length from h to this node), it includes the previous hop of the message in Phji.If i receives a request with fc < minfchji (meaning the request traversed a shorter path from h to this node), it resets minfchji to fc and Phji to the previous hop of the message. The set Phji contains the identifiers of nodes that can receive a corresponding RREP from i, if i sends one.When a destination node j receives a RREQ, it immediately sends back a RREP if fc ≤ minfchji. Every RREP explicitly specifies the set of nodes P that can accept it. The destination node j initializes this field to the previous hop of the RREQ, effectively indicating that the RREP is only intended for this node. Every RREP also has a hop count field hc initialized to zero by the destination.A node i processes a RREP if i∈ P. Node i then accepts the route in RREP if hc ≤ Dij or its existing routes to j have not been used for more than RouteF reshTime seconds and provided that the number of routes to j is less than or equal to MaxRoutes. If the received route is shorter (hc < Dij) or existing routes to j have not been used for more than RouteF reshTime seconds, Si j is reset to contain the previous hop of RREP. Node i then computes its distance Dij ← hc and forwards the message to its upstream nodes by setting P← Phji and incrementing hc by one if the corresponding request has not been replied yet. This process is repeated until the RREP reaches the source h.

Route Maintenance

Nodes rely on data packet acknowledgment provided by the link layer to determine the state of a link. Since packets are forwarded in a round-robin fashion, all links are periodically refreshed. Route maintenance occurs only when a node i loses all its active routes to some destination j after a data forwarding failure.

A link (i, k) is declared as ‘‘down’’ when node i does not receive an acknowledgment from the next hop node k after forwarding a data packet to k. When this occurs, k is deleted from Sij. If i has another route to j, it forwards all undeliverable packets through this route. If i has no other route to j, i broadcasts a RERR containing the header information of all data packets(excluding packets that are originating from this node) that cannot be delivered as a result of the link failure. If there are undeliverable packets originating from this node, i performs a route discovery.

Recall that it is possible for the RERR to contain header information of one or more data packets. Before parsing the RERR received from k, i creates a new RERR message that it will propagate upstream should it fail to salvage any packet. For every packet referred to in the message, node i performs the following:

Node i deletes k from Sij, where j is the destination of the packet. If i originated the referred packet and it has no other route j, it initiates a new route discovery if there is none in progress.

If i has other routes to j and it has a copy of the data packet in its data cache, it forwards the data packet according to the data forwarding rule. If i has no other route to j and it has a copy of the data being referred to in its data cache, it removes the referred data packet from its data cache and adds the data packet header information in the RERR it created. If i does not have the packet in its data cache and it is the previous hop of the packet, i adds the data packet header information in the RERR it created. If after parsing the RERR node i fails to salvage one or more data packets, it broadcasts the RERR it created.

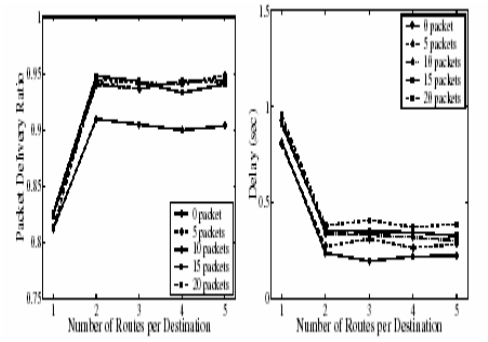
ALLOCATION GRANULARITY

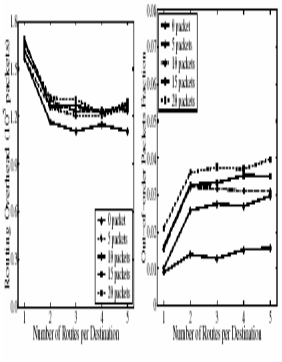
Here we use per packet allocation scheme, Two specifications,

.The first part evaluates the effect of data cache size and number of stored routes per destination on the performance of SMR

2. The second part is a performance comparison between SMR(using the optimum parameters derived in the first part), AODV and DSR. The network used for the simulations consists of 100 nodes in a 1500m × 600 m area. The movement of the nodes follows the ‘‘random way- point’’ model [9]. Six different pause times are used: 0, 30, 60,

120, 300, and 600 seconds. A zero pause time indicates that nodes are continuously moving while a 600-second pause time means that nodes are at rest for the entire simulation duration. Nodes move at speeds between 0 and 20 m/s.For simulations using constant-bit-rate (CBR) sources, the packet size is set to512 bytes and sending rate is set to 4





Performance of SMR as a function of the number of routes per destination. Each line plot represents a particular value for the data cache size. The text in the legend refers to the size of the data cache. packets per second. For simulations using TCP Reno sources, the packet size is set to 1460 bytes and maximum window to 8.

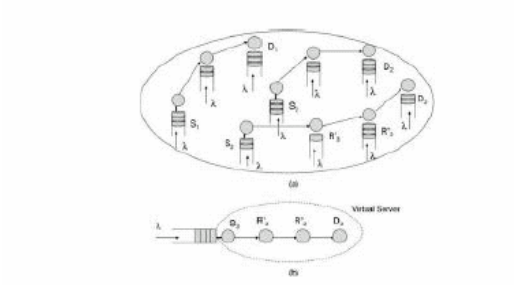


Fig Ad hoc wireless network with reservation-based multi-hop communications.

The main research directions considered in this work include (but are not limited to) the following.

1. We have first considered the case of reservation-based multi-hop communications among fixed nodes with regular topology: a pictorial scheme is shown in Figure This simple scenario has allowed to derive many insights about the behaviour of ad hoc wireless networks. For example, the concept of minimum spatial energy density, quantifying the minimum energy floating the network and necessary to guarantee communications, has been formalized, and closed-form expressions have been derived in various cases.A novel symbol-level approach to the evaluation of the inter-node interference affecting the communications has been derived.

2. We have introduced and the concept of effective transport capacity, which links the transport capacity, the information-theoretic quantity quantifying the rate distance product carried by the network, with the physical characteristics of the network. In particular, we have shown the existence of a drastic threshold between a sub-critical region(where connectivity rapidly drops to zero) and a supercritical region(where complete connectivity exists). Our results match with existing results based on the theory of continuum percolatio

3. Our approach has allowed to evaluate the impact of mobility through closed-for expressions, rather than through lengthy simulations.

4. Non-reservation-based ad hoc wireless networks have been considered, where a source-destination pair of nodes communicate through more than a single route

5. Unlike the conventional wisdom, we have shown that the concept of neighbors of a node is a good indicator of connectivity only if the topology is regular .

6. Topology of the nodes plays a major role in determining the network performance. We have evaluated the impact of node clusters , and we have clarified the relationship, in terms of obtainable performance,

between the cases of uniform and non-uniform topologies.

7. We have clarified the trade-off between transmitted power and connectivity: our results show that, for a given traffic load offered to the network, there exists an "optimal" value of the transmission power which a node should use.

8. Through a simple discrete-time communication model, the existence of an "optimal" channel utilization ratio, in order for the effective transport capacity of the network to be maximized, has been shown.

Routing Process

The routing process determines the paths between nodes in the network. It's the routing protocols function to control these events. This is a very complicated matter due to large involvement of all the nodes in the network. There are a several different factors to take into consideration when determining the paths between the nodes. One of the most challenging is the dynamic nature of the network. It's hard to keep track of the all the nodes whereabouts at all times. As a result of this, a reactive method has been developed to solve this issue. Unlike the proactive method the reactive doesn't need to know the nodes location at all times. Instead it only needs to make a request for a path when it's needed. The proactive method is commonly used in static networks or networks with a very low rate of mobility.

Classification of Routing Protocols

Most protocols can be classified in several ways. Some are classified as reactive or on- demand while others are proactive. In general, a proactive protocol finds routes in advance while a reactive protocol finds routes to the destination only when it absolutely must. For example, Ad hoc On demand Distance Vector routing (AODV) [AODV] is an on-demand protocol since no protocol information is transmitted before an application decides to send data and no data is sent until a route is formed, whereas Destination Sequenced Distance Vector protocol (DSDV) [DSDV] is a more proactive protocol in which routes are discovered and stored even before they are needed.

Proactive protocols generally generate much more traffic than on- demand protocols. A third general category is a hybrid algorithm that effectively combines multiple characteristics in a unique and meaningful way. For example, the Zone Routing Protocol (ZRP) [ZRP] is a hybrid protocol that combines local proactive routing with a globally reactive routing strategy.

PROBLEM SOLUTION

Multi path Routing

The use of multipath routing for ad hoc networks is not new and has been studied by several authors as extensions to existing protocols as well as for entirely new ones.If multipath routes are stored but only one path is used at a time, multipath routing is generally called alternate path routing, where as if more than one path is used at the same time it is referred to as simultaneous or disjoint multipath routing. Stojmenovic et al [Location] show via simulation that, while multipath routing may increase routing overhead while finding multiple routes, they have the potential for helping in network traffic load balancing, if data is sent simultaneously along multiple paths. In simulation studies on Ad hoc On demand Multipath Distance Vector Routing (AOMDV) [AOMDV], where data is sent via just a single path at a time, the authors stated that AOMDV, a multipath variant of AODV, improved the packet delivery ratios for CBR/UDP traffic by up to 40% and significantly reduced the packet delivery latency often by a factor of almost two.

They also stated that routing overhead in their method was improved by 30% since less route discovery phases were required versus AODV. They do note, however that at higher mobility the performance difference between AODV and AOMDV is much lower. In our experiments we find AOMDV does better than AODV in terms of delivery ratio in most scenarios at a cost of increased flooding. The AOMDV protocol is explained in more detail in Chapter 6.

The original Dynamic Source Routing (DSR) protocol includes the optimization of using an

alternate cached path when a path fails as an optimization, but did not explore it.

Some other authors have proposed multipath DSR and alternate path DSR protocols and evaluated their performance via simulation. Thus we were curious to study how well a position based algorithm which we call MLAR for Multipath LAR routes using a multipath route caching strategy versus other state of the art non-position based algorithms AODV (single path) and AOMDV (multi-path), as well as the position based algorithm LAR (single-path) in both 3D and 2D.

Multipath Routing Components

Multipath routing consists of three components: route discovery, route maintenance, and traffic allocation. We discuss these components in the following subsections.

Route Discovery and Maintenance

Route discovery and route maintenance consists of finding multiple routes between a source and destination node. Multipath routing protocols can attempt to find node disjoint, link disjoint, or non-disjoint routes. Node disjoint routes, also known as totally disjoint routes, have no nodes or links in common. Link disjoint routes have no links in common, but may have nodes in common. Non-disjoint routes can have nodes and link in common.

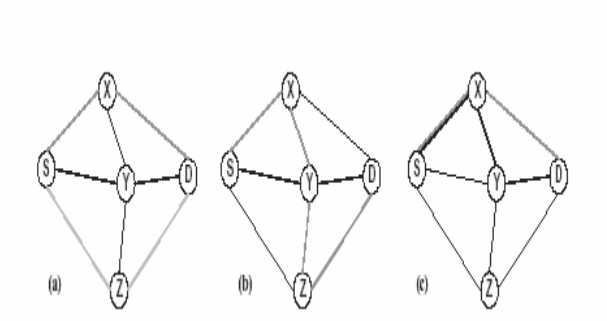


Fig (a) Routes SXD, SYD, and SZD in (a) have no links or nodes in common and are therefore node disjoint. Routes SXYZD and SYD in (b) have node Y in common and are therefore only link disjoint. Routes SXD and SXYD in (c) have node X and link SX in common and are therefore non-disjoint.

Disjoint routes offer certain advantages over non-disjoint routes. For instance, non-disjoint routes may have lower aggregate resources than disjoint routes, because non-disjoint routes share links or nodes. In principle, node disjoint routes offer the most aggregate resources, because neither links nor nodes are shared between the paths. Disjoint routes also provide higher fault-tolerance. When using non-disjoint routes, a single link or node failure can cause multiple routes to fail. In node or link disjoint routes, a link failure will only cause a single route to fail.

However, with link disjoint routes, a node failure can cause multiple routes that share that node to fail. Thus, node disjoint routes offer the highest degree of fault-tolerance.

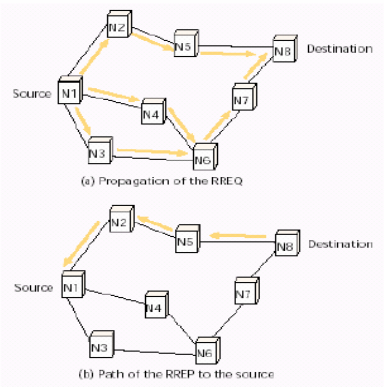


Fig (b,c) Route Discovery

Multipath Dynamic Source Routing

Multipath Dynamic Source Routing (MDSR) protocol proposed by A. Nasipuri and S.R. Das [NAS1] is the multipath extension to DSR. The basic idea is that when multiple flooded query messages arrive at the destination, apart from replying the query with the shortest route i.e. the primary route,The destination will also compute those source routes that are link- wise disjoint from the primary route. Disjoint routes are chosen so that a link failure in one route does not affect the others. When a route failure occurs in the primary route, alternate route will be used until a new route discovery initiated when all routes break down. The authors explored two variants, one where the

source gets multiple routes and another where all intermediate nodes on the primary route get multiple alternate routes.

First, alternate routes are only assigned to the source, then failure in intermediate link sends error packet back to use alternate routes causing a temporary loss of route for data packets. Improvement can be applied by equipping all intermediate nodes with a disjoint alternate route. Destinations need to replies to each intermediate node in the primary route with an alternate disjoint route to it. When an intermediate node encounters a transmission failure to the next hop, it may use alternate path to destination immediately instead of sending back error packet to source. Thus, only loss of both routes in a node generates an error packet back to the source. Intermediate node with alternate route to destination will stop the error packet and modifies source route on all later data packets to direct to its alternate route. The procedure continues until no alternate route along the primary route is available at all, a route discovery initiated.

The advantage of MDSR like MLAR is that it provides alternate paths for all intermediate nodes along the primary route.

The main disadvantage of MDSR is that this scheme will result in more route reply message flooding in the network, overhead for intermediate nodes' cache storing and computation overhead for the destination, particularly for the computing of alternate path of all the intermediate nodes.

The authors also found that multipath routing decreased the routing load but increased end to end delay as alternate routes tend to be longer than primary routes in their analytic results. They conclude that in a real network, a lower routing load would mean less interference and potentially lower end to end delay as well. The authors also found that the benefits of having more than 2 routes were minimal if any.

SECURITY THREATS

Dropping of Packets

These attacks are carried out by dropping either data packets or control packets. In the former case, a malicious node, after advertising a correct path to the destination, simply drops data packets to perform a denialof- service attack. In the latter case, a malicious node drops control packets sent by other nodes, while at the same time initiating a route discovery whenever it needs to send data. This is also known as routing misbehavior .

Fabrication Attacks

A malicious node can launch a denialof-service attack against a destination node by impersonating a node along a path to that destination and continually sending route error messages for the destination node. On receiving the route error messages, the nodes using that route will delete the route table entry for the destination node. In the absence of alternative routes, or by continually transmitting error messages, communication to the destination node can be successfully prevented.

Resource Depletion Attacks

Network bandwidth utilization is an important issue in a wireless ad hoc network. A single node or a group of nodes in collusion can perform a resource depletion attack. For example, two nodes can transfer large volumes of data between themselves, thereby clogging the network and depleting the available network bandwidth. A resource depletion attack can also be achieved by using control messages. For example, a malicious node can flood the network with route requests to random addresses. This would lead to an RREQ storm that depletes the available network bandwidth.

False Message Propagation Attacks

One instance of this type of attack is the redirection with modified sequence numbers. In this attack, a malicious node advertises a route to a node with a destination sequence number greater than the authentic value.By doing this, it diverts the traffic towards the attacker because the nodes will select the RREP with the highest destination sequence number. This attack is specific to AODV. Modification of the hopcount also has a similar effect on the routes chosen.

Misrouting Attacks

In this class of attacks, a malicious node attempts to send a data packet to the wrong destination. For example, this can be achieved by forwarding a data packet to the wrong next hop in the route to the destination or by modifying the final destination address of the data packet.

Man-in-the-Middle Attacks

A malicious node can combine the spoofing attack and the dropping of packets to perform a man-in-the-middle attack. The malicious node uses its place on the route as the first step in a man-in-the-middle attack by disallowing route requests to the destination. The attacker must be the only node within the range of the destination, or must be able to prevent any other route requests to the destination. The attacker sends a spoofed route reply to the source node and establishes a route with the source.

The attacker then sends a new route request to the destination node, establishes a route with the destination, and then drops the reply packet on its way to the source node. By doing this, the attacker controls the communication between the source and the destination.

Confidentiality and Privacy Attacks

Privacy guarantees non-disclosure of personal information stored at a node to any other node in the network. Con- fidentiality ensures that certain information is disclosed only to authorized entities.

CONCLUSION

In this paper we introduced cooperative packet caching, a strategy similar to cooperative caching aimed at reducing packet loss due to frequent route breakdowns. We have shown that because of the property of temporal locality in dropped packets, a small data cache (five packets) is sufficient to improve packet delivery. In essence, cooperative packet caching enables distributed packet salvaging. Although caching is now employed in many distributed systems, this is the first application in the network layer, particularly of ad hoc networks.

REFERENCES

[1] J. Broch, D. Maltz, D. Johnson, Y.C. Hu, and J. Jetcheva, ‘‘A performance comparison of multi-hop wireless ad hoc network routing protocols,’’ in Proc. ACM/IEEE MOBICOM ’98, Oct. 1998, pp. 85--97.

[2] M.S. Corson and J. Macker, ‘‘Rfc 2501: ad hoc networking (manet): Routing protocol performance issues and evaluation considerations,’’ Jan. 1999.

[3] M.S. Corson and A. Ephremides, ‘‘A distributed routing algorithm for wireless networks,’’ ACM/Baltzer Wireless Networks J., vol. 1, no. 1, pp. 61--82, Feb. 1995.

[4] R. Dube, C. Rais, K.Y. Wang, and S.K. Tripathi, ‘‘Signal stability- based adaptive routing (ssa)for ad hoc networks,’’ IEEE Personal Commun., vol. 4, no. 1, pp. 36--45, Feb. 1997.

[5] Z. Haas, ‘‘A new routing protocol for the recongurable wireless networks,’’ in Proc. IEEE ICUPC ’97, Oct. 1997.

[6] D. Johnson and D. Maltz, ‘‘Dynamic source routing in ad hoc wireless networks,’’ in Computing, T. Imielinski and H. Korth, Eds., pp.153--181. Kluwer Academic Publishers, Norwell, Mass., 1996.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

ANALYSING RISKS ON SOFTWARE TESTING

[1]Gowdhami.D, [2]ArunaDevi.P,

[1]Research Scholar, Dr SNS Rajalakshmi College of Arts and Science, Coimbatore

[2]Assistant professor, Dr SNS Rajalakshmi College of Arts and Science, Coimbatore

ABSTRACT

This paper speaks about the software testing phase in software development life cycle (SDLC) and types of software testing especially about the risk based software testing. There are several methods to risk based testing.In this Paper we are going to see the Fundamentals of risk analysis in software testing and also it will cover the types of risks in software testing and Benefits of risk based testing.

I. INTRODUCTION

Testing is the important phase in the software development life cycle. Testing will done after completing the development phase. Testing is the process of evaluating a system or its module(s) with the intent to find that whether it satisfies the specified requirements or not. This activity results in the actual, expected and difference between their results. A most appropriate definition of testing id defined below.

Testing is the process of executing a program with the intent of finding errors

II. WHY TESTING IS NECESSARY

Testing is necessary because we all make mistakes. Some of those mistakes are unimportant, but some of them are expensive or dangerous. We need to check each and everything we produce because things can always go wrong

III. TESTING OBJECTIVE AND PURPOSE

Testing has different goals and objectives:

- Finding defects
- Gaining confidence in and providing information about the level of quality.
- Preventing defects

IV. TYPES OF SOFTWARE TESTING

Few types of software testing are explained below

A. *Black box testing* – It is otherwise called input/output- driven testing. Internal system design is not considered in this type of testing. Tests are based on requirements and functionality.

B. *White box testing* – It is otherwise called Logic –driven Testing. This testing is based on knowledge of the internal logic of an application's code. Also known as Glass box Testing. Internal software and code working should be known for this type of testing. Tests are based on coverage of code statements, branches, paths, conditions.

C. *Unit testing* – Testing of individual software components or modules. Typically done by the programmer and not by testers, as it requires detailed knowledge of the internal program design and code. May require developing test driver modules or test binds.

D. *Integrationtesting* –It means a testing of several modules in order to verify that the interfaces are defined and implemented correctly. It is orderly progression of testing in which software components and hardware components are combined in highly controlled and

orderly manner and tested until the entire project has been developed.

E. *Functional testing* – This type of testing ignores the internal parts and focus on the output is as per requirement or not. Black-box type testing geared to functional requirements of an application.

F. *System testing* – It is the process of exercising the integrated software system with the objective of finding discrepancies between the performance of a system and its original purposes.

G. *End-to-end testing* – Similar to system testing, involves testing of a complete application environment in a situation that mimics real-world use, such as interacting with a database, using network communications, or interacting with other hardware, applications, or systems if appropriate.

H. *Stress testing* – System is stressed beyond its specifications to check how and when it fails. Performed under heavy load like putting large number beyond storage capacity, complex database queries, continuous input to system or database load.

I. *Performance testing* – Term often used interchangeably with 'stress' and 'load' testing.To check whether system meets performance requirements. Used different performance and load tools to do this.

J. *Risk Based Testing* **-** Risk based testing uses risk to prioritize and emphasize the appropriate tests during test execution.

V. RISK BASED TESTING

What Is Risk?

Risk is the possibility of a negative or undesirable outcome or event.Some possibilities of risks meet in software projects are defined below.

A. Technicalrisks include problems with languages, project size, project functionality, platforms, methods, standards, or processes. These risks may result from unnecessary constraints, lack of experience, poorly defined parameters, or dependencies on organizations outside the direct control of the project team.

B. Management risks include lack of planning, lack of management experience and training, communications problems, organizational issues, lack of authority, and control problems.

C. Financial risks include cash flow, capital and budgetary issues, and return on investment constraints.

D. Contractual and legal risks include changing requirements, market-driven schedules, health & safety issues, government regulation, and product warranty issues.

E. Personnel risks include staffing lags, experience and training problems, ethical and moral issues, staff conflicts, and productivity issues.

F. Other resource risks include unavailability or late delivery of equipment & supplies, inadequate tools, inadequate facilities, distributed locations, unavailability of computer resources, and slow response times.

VI. SOFTWARE RISK

Software risk can be classified into 2 different kinds. They are

- Project Risk
- Product Risk

A. Project Risk:

The risks associated with the testing activity which can endanger the test project cycle is known as project risk.

Some of the risks associated with project are:

- Delay in the test build to test team.
- Unavailability of test environment.
- Delay in fixing test environment due to lack of system management.
- Delay in fixing faults by development team.
- Organizational problems which can be like shortage of man power.Skill power etc.
- Major changes in the SRS which invalidates the test cases and requires changes in the test case.

B. Product Risk

Product risk is the possibility that the system or software might fail to satisfy or fulfill some reasonable expectation of the customer, user, or stakeholder.

The product risks that can put the product or software in danger are:

- If the software skips some key function that the customers specified, the users required or the stakeholders were promised.
- If the software is unreliable and frequently fails to work.
- If software fail in ways that cause financial or some other damage to a user or the company that user works for.
- If the software has problems related to a particular quality characteristic, which might not be functionality, but rather security, reliability, usability, maintainability or performance.

VII. ADVANTAGES OF RISK-BASED TESTING:

- Improved quality – All of the critical functions of the application are tested. Real time clear understanding of project risk.
- Give more attention on risks of the business project instead of the functionality of the information system.
- Provides a negotiating instrument to client and test manager similar when existing means are limited.
- Associate the product risk to the requirement identifies gaps.
- During testing, test reporting always takes place in a language (risks) that all stake-holder understands.
- Testing always concentrate on the most important matters firstwith optimal test delivery, in case of – limited time, money and qualified resources. With the time and resources we have, we just can able to complete 100% testing, so we needto determine a better way to rush our testing effort with still managing the risk of the application under test. Efforts are not wasted on non-critical or low risk functions.
- Improve customer satisfaction – Due to customer involvement and good reporting and progress tracking.

VIII. RISK ANALYSIS FUNDAMENTALS IN SOFTWARE TESTING

A. Risk Identification

Risk identification is the first step in risk management. We need to identify both project and product risk by using certain techniques. Some of the most common techniques which can be applied to identify different risks are using risk templates, interviewing the stakeholders, project demonstrations etc.

Several formal techniques like Failure Mode and Effect Analysis (FMEA) and Failure Mode Effect and Criticality Analysis (FMECA) are used to find the risk. These techniques identify the effects of the risk if in case that becomes an outcome. The effects can be on people, society, users, customers etc.

B. Risk Strategy

Risk based strategizing and planning involves the identification and assessment of risks and the development of contingency plans for possible alternative project activity or the mitigation of all risks. These plans are then used to direct the management of risks during the software testing activities. It is therefore possible to define an appropriate level of testing per function based on the risk assessment of the function. This approach also allows for additional testing to be defined for functions that are critical or are identified as high risk as a result of testing (due to poor design, quality, documentation, etc.).

C. Risk Assessment

Assessing risks means determining the effects (including costs) of potential risks. Risk assessments involve asking questions such as: Is this a risk or not? How serious is the risk? What are the consequences? What is the likelihood of this risk happening? Decisions are made based on the risk being assessed. The decision(s) may be to mitigate, manage or ignore.

Three perspectives of Risk Assessment

- *Effect***:** To assess risk by Effect, identify a condition, event or action and try to determine its impact.

- *Cause***:** To asses risk by Cause is opposite of by Effect. Begin by stating an undesirable event or condition and identify the set of events that could have permitted the condition to exist.

- *Likelihood***:** To asses risk by Likelihood is to determine the probability that a requirement will not be satisfied.

D. Risk Mitigation

The activity of mitigating and avoiding risks is based on information gained from the previous activities of identifying, planning, and assessing risks. Risk mitigation/avoidance activities avoid risks or minimise their impact.

The idea is to use inspection and/or focus testing on the critical functions to minimise the impact a failure in this function will have in production.

E. Risk Reporting

Risk reporting is based on information obtained from the previous topics (those of identifying, planning, assessing, and mitigating risks).

In the test phase it is important to monitor the number of errors found, number of errors per function, classification of errors, number of hours testing per error, number of hours in fixing per errors etc.

F. Risk Prediction

Risk prediction is derived from the previous activities of identifying, planning, assessing, mitigating, and reporting risks. Risk prediction involves forecasting risks using the history and knowledge of previously identified risks.

During test execution it is important to monitor the quality of each individual function (number of faults found), and to add additional testing or even reject the function and send it back to development if the quality is unacceptable. This is an on-going activity throughout the test phase.

IX. RISK ANALYSIS DURING SOFTWARE TESTING

Some of the Common risks could be:

- New Hardware.
- New Technology.
- New Automation Tool.
- Sequence of code delivery.
- Lack of test resources.

In Software Testing some unavoidable risk might takes place like:

- Change in requirements or incomplete requirements.
- Time allocation for testing.
- Developers delaying to deliver the build for testing.
- Urgency from client for delivery.

To overcome these risks, the following activities can be done.

- Conducting Risk Assessment review meeting with the development team.
- Profile for Risk coverage is created by mentioning the importance of each area.
- Using maximum resources to work on High Risk areas like allocating more testers for

High risk areas and minimum resources for Medium and Low risk areas.

- Creation of Risk assessment database for future maintenance and management review.
- Identify and describe the risk magnitude indicators: High, Medium and Low

X. CONCLUSION

In this paper we have presented the types of software testing and also about the risks on the software testing. We have analysed the risks on software testing using risk identification, risk strategy, risk assessment, risk mitigation, risk reporting and risk prediction.

References:

[1] Roger Pressman.S., " Software Engineering : A Practitioner's Approach ", 3rdEdition, McGraw Hill, 1997.

[2] Boris Beizer," Software Testing Techniques", 2nd Edition, Dreamtech Press, 2003

[3] Marnie.L. Hutcheson, "Software Testing Fundamentals", Wiley, 2007

[4] Software Quality Producing Practical, Consistent Software By Mordechai Ben-Menachem and Garry S.Marliss

[5] The Art of Software Testing, Second Edition By Glenford J. Myers Revised and Updated by Tom Badgett and Todd M. Thomas withCorey Sandler

[6] http://istqbexamcertification.com

[7] http://www.softwaretestingmentor.com

[8] http://globaltesters.wordpress.com

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Deblurring of image using Noise suppressed steering kernel

Janet Lobo
Providence College for Women,
Springfield, the Nilgiris
janetlobosjt@gmail.com

ABSTRACT

Debluring is the process of eliminating the noise and blur in the image. There are number of processes that are used to remove the blur in the image. The deblurring techniques give the maximum of best result when the fuzzy technique is applied. .In digital image restoration regression is one of the processes of recovering original image which is affected by noise or blur. The image can be recovered by using denoising and deblurring techniques. The steering kernel makes use of the few steps iteratively and gives the best result.

Key words: Blurring, deblurring, steering kernel.

I. INTRODUCTION

Image processing is the process that people use much in the field of enhancing the original images. People need it to make the image clean and clear. Especially the images that are retrieved from the satellite and the images that are transferred over the internet reach the destination with noise. Blur and noise are the main two problems that are found in the Digital image processing. During important camera setting if the exposure time is fixed, a large crack will increase the noise ratio and out of focus blur. This removes the high frequency components of the image. The hole even if it is small will increase the noise and the blur in the image. Image processing is the technique used to enhance the images received from the cameras, satellite and space. Restoring an image is a process that is accepted mostly to recover the original image.

II. BLURRING

Blur in an image is caused by camera or object movement or by not properly focusing the object. The causes for the blur are moving the camera during image capture process or when the long time is spent by the subject, use of wide angle lens or scattered light alteration.[1][2]

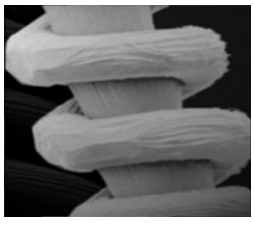


Fig 1. Blur image

1. Blurring Types

In digital image processing there are 3 common types of Blurring effects:

- Average blur
- Motion blur
- Gaussian blur

1.1 Average blur

The average blur is used to remove noise and speckles from an image. It is used when the noise is present in the entire image. This type of blurring can be distribution in horizontal and vertical direction and can be circular averaging by radius R which evaluated by the formula:

$R = \sqrt{h2 + v2}$

Where: h is the horizontal size blurring direction and v is vertical blurring size direction of the average blurring.

1.2 Motion blur

The Motion Blur is a filter that makes the image appear to be moving by adding a blur in a specific direction. The motion can be controlled by angle or direction (0 to 360 degrees or –90 to +90) and/or by distance or intensity in pixels based on the software used. [5]

1.3 Gaussian blur

The Gaussian Blur effect is a filter that blends a specific number of pixels incrementally, following a bell-shaped curve. The blurring is dense in the center and feathers at the edge. Apply

Gaussian Blur to an image when you want more control over the Blur effect.

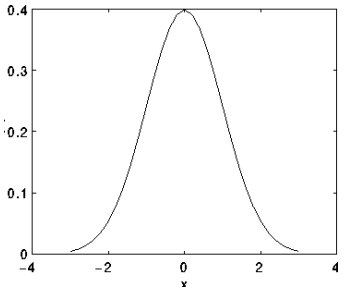


Fig 2. Gaussian blur

2. Methods of Deblurring

2.1 Weiner filtering

Wiener filter is a method of restoring image in the presence of blur and noise. The Wiener filtering minimizes the overall mean square error in the process of inverse filtering and noise smoothing. The Wiener filtering is a linear estimation of the original image.Wiener filter has two separate part, an inverse filtering and a noise smoothing. It not only performs the deconvolution by inverse filtering (highpass filtering) but also removes the noise with a lowpass.The disadvantage of this implementation is that when the inverse filter is singular, we have to use the generalized inverse filtering.

2.2 Inverse Filtering

Inverse filtering is the easiest way to restore the image. Since the inverse filter is a high pass filter it returns the noise that is present in the image.

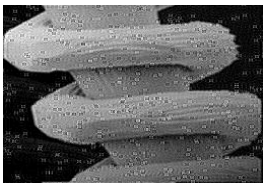


Fig 3. Inverse filtering

There are two types of inverse filters.

- Thresholding method
- Iterative method

Thresholding method

A blurred image can be modeled as

$$g(n_1, n_2) = f(n_1, n_2) * *b(n_1, n_2)$$

where *f* is the original image, *b* is some kind of a low pass filter and *g* is our blurred image. To find the original image one has to change the blur function into another type of high pass filter.

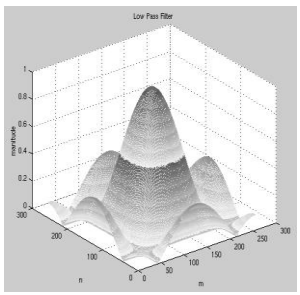


Fig 4.Thresholding

In inverse filtering the noise remains the same and not many changes takes place. Since it is a high pass filter it does not give a good result.Here Edge sharpness is increased but the noise also is increased. If *n* is increased the noise can be reduced but the edges will not be sharp as before.

Iterative method

The idea behind the iterative procedure is to make some initial guess of f based on *g* and to update that guess after every iteration. The procedure is

$$\hat{f}_0(n_1, n_2) = \lambda g(n_1, n_2)$$

$$\hat{f}_{k+1}(n_1, n_2) = \hat{f}_k(n_1, n_2) + \lambda(g(n_1, n_2) - \hat{f}_k(n_1, n_2) * *b(n_1, n_2))$$

where $\hat{f}_0$ is an initial guess based on *g*. If our $\hat{f}_k$ is a good guess, eventually $\hat{f}_k$ convolved with *b* will be close to *g*. When that happens the second term in the $\hat{f}_{k+1}$ equation will disappear and $\hat{f}_k$ and $\hat{f}_{k+1}$ will converge. λisthe convergence factor and it lets us determine how fast $\hat{f}_k$ and $\hat{f}_{k+1}$ converge.

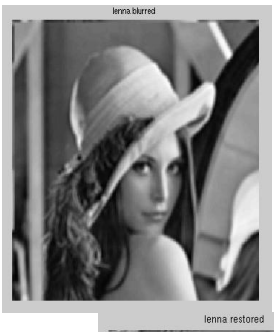




Fig 5 original image Fig 6 restored image

Generally the iterative method will not get the exact result like that of inverse filtering, rather it produces less noise.

The figure () is noise free after the 150 iterations.

2.3 Wavelet-based Image Restoration

The wavelet based transformation improves the performance and gives a better result. In Wavelet based deconvolution the Wiener filter is modified and applied in wavelet domain. The regularized inverse filter changes the order of regularized inverse filter and wavelet transform. This can be used only if wavelet transform and blurring function is separable. For interpretation the order can be exchanged for blurring operation and wavelet transform.[3]

2.4 Lucy-Richardson method

The Richardson–Lucy algorithm, also known as Richardson–Lucy deconvolution, is an iterative procedure for recovering a hidden**image** that has been the blurred.



Fig 7. Lucy Richardson filter

2.5 Blind deconvolutionmethod

Definition of the blind deblurring method can be expressed by:

$g(x, y) = PSF * f(x,y) + \eta(x,y)$



Fig 8.blind deblur Image

Where: g (x, y) is the observed image, PSF is Point Spread Function, f (x,y) is the constructed image and η(x,y) is the additive noise term.

2.6 Gaussian Deblurring

The Wiener filter is used for deblurring an image in the case when the blur kernel (point spread function) is known. The computation of the Fourier transform of the deblurred image F by F = (G/H) *(|H|^2/ (k+|H|^2)) where

G = Fourier transform of original blurry image

H = Fourier transform of blur kernel

k = deblurring parameter ($k \geq 0$)

Setting the value of k is the hard part. The value should depend on the amount of noiseexpected in the image. If the k value is tried repeatedly the size of the image can be matched[4].

III. Noise suppressed steering kernel for Image Deblurring

Traditionally the steering kernel technique is a non parametric approach. It depends on data

itself rather than depending on any specific model. The operations of steering kernel like calculating local gradients, assigning weights consumes more time. So the parallelism for computation is used to reduce this problem. When the iterative process is continued the unauthentic edges are added and they become stronger. This problem can be overcome by using the median filters at the end of each iteration, since this method is an iterative process.

Steps of steering kernel

1. Importing image
2. Formulating image and local gradient.
3. Parametric analysis
4. Steering matrix formation
5. Application of steering kernel
6. Suppression of spurious edges
7. Thresholding

1. Importing the Image

The image that needs to be processed is imported from the database. All the processes that need to be performed are done initially. The initial processes are enhancing quality, cropping and resizing the image.

2. Formulating the image and local gradient

The image gradient is the directional change in the intensity or color in the image. This is used to extract the information from the images. This is one of the most important parameter used in the image processing.

3. Parametric analysis

The different parameters that are used here are scaling parameters, rotational parameters and elongation parameters. Scaling parameters provide information about the size of the image. Rotational parameters which define on what angle the image need to be located or rotated. The elongation parameters are those that imply how the image has been elongated.

4. Steering matrix formation

The steering kernel regression is carried out using the steering kernel. The data not only depends on sample location and density but also on the radiometric properties. The regression is adapted locally so that the kernels arrange themselves along the features instead of arranging themselves in across.

5. Suppression of spurious edges

The image after performing all the above steps contains some noise in it. When this output image is subjected to iterations it will keep on increase the noise content. So in this proposed method the noise content is removed before letting the image into next iterations so that the noise won't get added up. The noise i.e. the spurious edges is removed by passing the image into a median filter and then allowed for iterations.

6. Thresholding

After the spurious edges are being removed the output image is subjected to a process of thresholding. The output values are being compared with one defined value which is the threshold value such that the output image values are compared with this defined value. If the appropriate threshold condition is not met the whole process will repeat from the initial stage of taking gradient. If it is met the obtained output is the final de-blurred output.

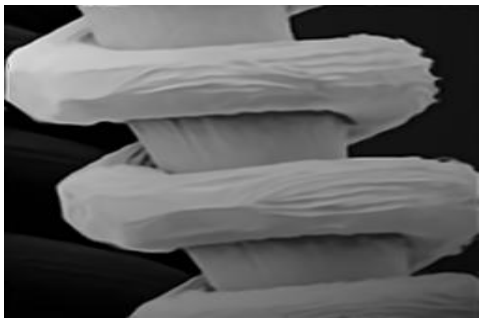


Fig 9 resulting image

This image is the result of applying the median filter iteratively along with steering kernel. The steps that are provided by the steering Kernel make the image run iteratively till the best result is found.

IV. Conclusion

The steps that are used in the Steering Kernel produce the best result. The clear image can be retrieved and all the noise and the blurs can be removed through the iterative processes applied along with the median filter. This process can be used anywhere and all who deal with photography.

The Image Processing has lot more to learn in clearing the images and bringing it to 100% clarity.

References

[1] J.Harikiran, B.Saichandana, B.Divakar Noise Removal in Digital Images. International journal Volume 10– No.8, 39-42.

[2] Fundamentals Of Digital Image Processing- Anil K. Jain, Prentice-Hall, 1989

[3] Charles Boncelet (2005).”Image Noise Models”. In Alan C. Bovik. Handbook of Image and Video Processing.

[4] Gonzalez and R. Woods *Digital Image Processing*, Addison-Wesley Publishing Company, 1992, p 191.

[5] Joao P.A. Oliveira, Mario A.T.F igueiredo, and Jose M. Bioucas-Dias, Blind Estimation of Motion Blur Parameters for Image Deconvolution, Fundacao Para a Ciencia Technologic, under project POSC/EEA-CPS/61271/2004.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

PERFORMANCE OF MULTI SERVER AUTHENTICATION AND KEY AGREEMENT IN NETWORK SECURITY

[1]Ms. S.Gowthami [2]Ms.R.Saveetha
[12]Assistant Professor, Kongu Arts and Science College Erode

Abstract

Using smart cards, remote user authentication and key agreement can be simplified, flexible, and efficient for creating a secure distributed computers environment. Addition to user authentication and key distribution, it is very useful for providing identity privacy for users. In this paper, we propose novel multi server authentication and key agreement schemes with user protection in network security. We first propose a single-server scheme and then apply this scheme to a multi-server environment. The main merits include:(1) The privacy of users can be ensured; (2) a user can freely choose his own password; (3) the computation and communication cost is very low; (4) servers and users can authenticate each other; (5) it generates a session key agreed by the server and the user; (6) our proposed schemes are Nonce-based schemes which does not have a serious time synchronization problem.

Keywords: Network security, privacy protection, session key, smart card, user authentication

1. INTRODUCTION:

For obtaining permitted services by service providers in a network environment, the user must legally login to the provider's server. In general, the user transmits a message of user authentication to the server, and then the server must be able to verify the identity of the user and give him the right of using permitted services. Typically, the user passes a password as a secret token to the server. The server first checks if the user's identity and the password are matching. The server rejects the user's request if his Identity or the password is not matching. If the password is matching, the server give the user the right for using the permitted services. In 1981, Lamport first proposed a password
authentication scheme at the both ends of the communication. Since then, many schemes have been proposed to point out its drawback and improve the security and efficiency of Lamport's scheme. Only passing a password for authenticating between the user and the server is not enough, since it is less safety and is easily tapped by the adversary. Before two parties can do secure communication, a session key is needed for protecting subsequence communications. Also, using smart cards, remote user authentication and key agreement can be simplified, flexible and efficient for creating a secure distributed computers environment. It is also useful for providing identity privacy for the users.

In 2004, Juang proposed two efficient authentication and key agreement schemes for single server, and multi server environments. But both Juang's schemes have no ability of anonymity for the user. Yang et. al. Proposed user identification and key distribution scheme with the ability of privacy protection but we point out it is less efficient because of using public-key cryptosystems. For basically security and efficient requirements, the following criteria are important for remote user authentication and key agreement schemes with smart cards.

C1. Privacy protection : When the user authenticates successfully to the server, the adversary can not derive the user's identity.

C2. Freely chosen password: Users can freely chosen and change their passwords for protecting their smart cards.

C3. Low computation and communication cost: Since capacity and communication constrains of smart cards, they may not offer a powerful computation capability and high bandwidth.

C4. Mutual authentication: Servers and users can authenticate each other.

C5. Session key agreement: Servers and users must negotiate a session key for subsequent Communications.

2. AUTHENTICATION:

Authentication systems can be categorized according to the number of identification factors required to ascertain identity.

- Single-factor authentication uses user ID/password combinations to prove identity.
- Two-factor authentication requires two components, usually a combination of something the user knows

(Such as a password) and something the user possesses (such as a physical token Secure ID card).

- Three-factor authentication adds a biometric, a measurement of a human body characteristic.

The more authentication factors used, the more secure the process. However, the more factors you add, the more you add complexity, cost, and management overhead. Every scenario will offer a different break-even point in the trade-off between simplicity and security.

Single-factor authentication with user ID and password is the most common authentication system today. It's easy to administer, familiar to users, and can provide a high level of security if strong password procedures are enforced. Legacy password systems have had some challenges, however, since multiple strong passwords are very hard for users to remember. The recommendations in this section will show how this problem can be minimized with a "Single Strong Password" system.

Tokens such as smartcards and Secure ID cards are added as a second factor in many authentication systems—requiring that the user have physical possession of the token. An attacker would similarly have to have possession of the user's token in order to gain system access. The higher level of authentication comes with additional system cost, however, due to the necessary tokens and token readers. In addition, tokens can be easily lost, which can present a high administration overhead for reissuing. Biometric factors for authentication measure characteristics of the user's body such as fingerprint, handprint, retina, iris, or voice characteristics. Biometric measurements are a useful additional factor and add an even higher level of authentication security. A biometric authentication system entails a measurement proving whom the person actually is, rather than proving they have something such as a token or proving that they know something such as a password. Unfortunately, biometric measurements are not 100 percent effective; with the present state of the technology, it is possible to register false positives and false negatives. Biometric authentication systems also require biometric readers at system access points, adding new system costs. Strong cryptographically-based authentication can be provided through the use of digital certificates issued to users and stored on tokens or within the user's computer memory. Cryptographic algorithms are used to ensure that a particular certificate has been legitimately issued to the user. A Public Key Infrastructure is used to enable the issuance and maintenance of digital certificates. Strong cryptographically-based systems provide very stringent authentication. However, these systems are expensive and incur additional management overhead. Therefore, they are currently being adopted only in very secure environments.

2.1. Authorization:

Once authenticated, authorization mechanisms control user access to appropriate system resources. Authorization can be categorized according to the granularity of control; that is, according to how detailed a division is made between system resources. Fine-grained authorization refers generically to a system where access is controlled to very fine increments, such as to individual applications or services.

Authorization is often "role based" whereby access to system resources is based on a person's assigned role in an organization. The System Administrator role may have highly privileged access to all system resources whereas the General User role would only have access to a subset of these resources. Finer grained authorization can be applied to define other roles, such as a Human Resources Administrators role that has exclusive access to confidential HR databases, and an Accounting role that has exclusive access to accounting systems.

Authorization may also be "rules based" whereby access to system resources is based on specific rules associated with each user, independent of their role in the organization. For example, rules may be set up to allow Read Only access or Read/Write access all or certain files within a system, or access only during certain times or from certain devices.

2.2. Authentication and authorization protocols :

Several protocols have been commonly adopted for authentication services. The RADIUS protocol (Remote Authentication Dial In User Service – IETF RFC2865) is widely used to centralize password

authentication services. Originally designed to authenticate remote dial-in users, the RADIUS protocol has been adopted for general user authentication services. Recently, the LDAP (lightweight directory access protocol – IETF RFC2251) has been finding extensive use in authentication and authorization systems. LDAP provides a convenient method for storing user authentication and authorization credentials. RADIUS authentication servers are often coupled with credential storage in LDAP directories to provide centralized authentication and authorization. When a user attempts to access a particular application on such a system, the application queries the user for authentication credentials

• **Servers**, including departmental servers for user access and file sharing, central application servers such as finance and databases, and Web servers for either public Web or Intranet applications.
• **Operating systems**, typically multiple versions of multiple operating systems running on servers and clients.
• **Network devices**, including routers, Layer 4-7 load-balancing switches, Layer 3 core switches, Layer 2 distribution switches, and wireless LAN access points.
• **Security devices**, such as firewalls, VPN gateways, intrusion-detection and anti-virus servers, SSL accelerators, authentication servers, and content filtering servers.

3 Securing the data center:

The typical enterprise data center supports mission-critical applications and houses a high concentration of capital-intensive resources and confidential data—all connected to the inherently insecure Internet as well as internal users. That means securing the data center presents some unique requirements for failsafe security without compromising performance and availability for users. The need increases as enterprises discover new ways to exploit high-performance, Internet-empowered data centers:

• **Ensure business continuity.** Massive processing throughput and transport bandwidth now make it feasible to store primary and duplicate sets of critical data in multiple data centers, in real time—to extend business continuity services, real-time storage mirroring, and live backup across service provider networks.

Support critical business applications. Enterprises use data centers to host business applications, implement firewalls or virtual private networks, provide storage services and content delivery of static and streaming media, and more.

4. SINGLE SERVER AUTHENTICATION AND KEY AGREEMENT WITH USER ANONYMITY:

In this section, we propose an efficient single server user authentication and key agreement scheme with privacy protection. The concept used in this section will be used in the next section to construct an efficient multi-server user authentication and key agreement scheme with privacy protection. Let ID_i be a unique identification of user i. Also, let x be the master secret key kept secretly by the server S.

5.1. The Proposed Scheme

The proposed scheme is as follows.

5.1.1. Registration Phase:

Assume U_i submits his identity ID_i and his password PW_i to the server S for registration. If S accepts this request, he will perform the following steps:

Step 1: Compute U_i's secret information $\alpha_i = h(x||ID_i)$ and $\beta_i = \alpha_i$® PW_i. Compute the pseudo identification number $\lambda_{i,1} = h(\alpha_i ||ID_i||1)$ and records $(k = 1, \lambda_{i,1}, ID_i)$ in an identification table

Step 2: Store IDi, $\lambda_{i,1}$, k = 1, and β_i to the memory of a smart card and issue this smart card to U_i or send them secretly to U_i.

User Authentication and Session Key Agreement Phase:

If U_i wants to log into S anonymously, he must attach his smart card to a card reader. He then inputs his identity ID_i and his password PW_i to this device. The following protocol is the kth login with respect to this smart card.

Step 1: $U_i \rightarrow S : N_1, \lambda_{i,k}, E_{\alpha i} (ru_k, h(N_1||ru_k|| \lambda_{i,k}))$;

Step 2: $S \rightarrow U_i : N_2, E_{\alpha i} (rs_k, h(rs_k||N_1||N_2))$;

Step 3: $Ui \rightarrow S : E_{skk} (N_2 + 1)$.

. **Performance Considerations:**

We evaluate the efficiency of our scheme and Juang's scheme in Table 1. First, we assume the block size of secure symmetric cryptosystems is 128 bits and the output size of secure one way hashing functions is 128 bits. Because both our proposed single- server scheme and Juang's scheme are based on symmetric key cryptosystem, the performance is very well. In our scheme and [8], the password length only 128 bits is required. Our proposed scheme needs 384 bits for the user authentication. Both ours and Juang's scheme [8], the computation cost for registration is only needed one hash operation. The computation cost are aggregated operation numbers, including encryption operations, decryption operations or hashing operations. The encryption and encryption operations may be asymmetric or symmetric cryptosystem. In the login and session key agreement phase of our scheme, three symmetric key encryptions, three symmetric key decryptions and seven hash operations are required. In that of Juang's scheme [8], only three symmetric key encryptions, three symmetric key decryptions and three hash operation are required. The computation cost of the login and session key agreement is not including cost of generating session key. Although our proposed scheme has a little high communication and computation cost than Juang's scheme [8], but our scheme have more complete functionality. The functionality comparison between our proposed scheme and related scheme is given in Table 2. Compared

5.MULTI- SERVER AUTHENTICATIONS AND KEY AGREEMENT WITH USER ANONYMITY:

There are three kinds of participants in our multi-server protocol: a key distribution centre, service providers (servers) and users. Let KDC denote the trusted key distribution centre, U_i denote user i, S_j denote service provider j. Let UID_i be a unique identification of U_i and SID_j be a unique identification of service provider j. Also, let x be the master secret key kept secretly by the key distribution centre KDC and $\delta_j = h(x||SID_j)$ be the secret key shared by S_j and KDC. The shared secret key δ_j can be computed by KDC and sent secretly to S_j after he registered at KDC.

The Proposed Scheme:

The proposed scheme is as follows.

Registration Phase:

Assume Ui submits his identity UIDi and his password PWi to KDC for registration. If KDC accepts this request, he will perform the following steps:

Step 1: Compute Ui's secret information _i = h(x || UIDi) and _i = _i _ PWi.

Step 2: Store UIDi, and _i to the memory of a smart card and issue this smart card to Ui or send them secretly to Ui.

6. CONCLUSIONS:

In this paper, we have proposed two user authentication and key agreement schemes with privacy protection for single server and multi- server environments. Regarding the single-server scheme, it is more simple and efficient. Regarding the multi-server scheme, users only need to register one time and can use all provided services by service providers. Both our proposed schemes have the ability of privacy protection. Our schemes also have low communication and computation cost for user authentication by only using symmetric cryptosystems and one-way functions. Also, our schemes successfully solve the serious time-synchronization problem in a distributed computers environment since our proposed schemes are nonce-based.

7.REFERENCES:

[1]S. Bellovin and M. Merritt, "Encrypted key exchange: Password-based protocols secure against dictionary attacks," in Proceedings of IEEE Symposium on Research in Security and Privacy, pp. 72-84, 1992.

[2] M. Burrows, M. Abadi, and R. Needham, "A Logic of Authentication," ACM Transactions on Computer Systems, vol. 8, no. 1, pp. 18-36, 1990.

[3]Y. Chang and C. Chang, "Authentication schemes with no verification table," Applied Mathematics and Computation, vol. 167, pp. 820-832, 2005.

[4]W. Diffie and M. Hellman, "New directions in cryptography," IEEE Transactions on Information Theory, vol. IT-22, no. 6, pp. 644-654, 1976.

[5]C. Fan, Y. Chan, and Z. Zhang, "Robust remote authentication scheme with smart cards," Computers & Security, vol. 24, pp. 619-628, 2005.

[6]M. Hwang, C. Lee, and Y. Tang, "A simple remote user authentication scheme," Mathematical and Computer Modelling, vol. 36, pp. 103-107, 2002.

[7] T. Hwang and W. Ku, "Repairable key distribution

protocols for internet environments," IEEE Transactions on Communications, vol. 43, no. 5, pp. 1947-1950, 1995.

[8] W. Juang, "Efficient password authenticated key agreement using smart cards," Computers & Security, vol. 23, no. 2, pp. 167-173, 2004.

[9] W. Juang, "Efficient multi-server password authenticated key agreement using smart cards," IEEE Transactions on Consumer Electronics, vol. 50, no.1, pp. 251-255, 2004.

[10] W. Juang and W. Nien, "Efficient password authenticated key agreement using bilinear pairings," in the 16th Information Security Conference, pp. 214-221,Taichung, Taiwan, June 2006.

[11] L. Lamport, "Password authentication with insecure communication," Communications of the ACM, vol. 24, pp. 770-772, 1981.

[12] A. Lenstra, E. Tromer, A. Shamir, W. Kortsmit, B.Dodson, J. Hughes, and P. Leyland, "Factoring estimates for a 1024-bit RSA modulus," in Advances in Cryptology (Asiacrypt'03), LNCS 2894, pp. 55-74, Springer, New York, 2003.

[13]I. Lin, M. Hwang, and L. Li, "A new remote user authentication scheme for multi-server architecture," Future Generation Computer Systems, vol. 19, pp.13-22, 2003.

[14] W. Ku and S. Chen, "Weaknesses and improvements of an efficient password based remote user authentication scheme using smart cards," IEEE Transactions on Consumer Electronics, vol. 50, no. 1, pp. 204-207, 2004.

[15]R. Merkle, "One way hash functions and DES," in Advances in Cryptology (Crypt'89), LNCS 435, pp. 428-446, Springer, New York, 1989.

[16] NIST FIPS PUB 180-2, Secure Hash Standard, National Institute of Standards and Technology, U. S. Department of Commerce, DRAFT, 2004.

[17] NIST FIPS PUB 197, Announcing the Advanced Encryption Standard (AES), National Institute of Standards and Technology, U. S. Department of Commerce, Nov. 2001.

[18] D. Seo and P. Sweeney, "Simple authenticated key agreement algorithm, Electronics Letters, vol. 35, pp. 1073-1074, 1999.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Cloud Computing Security From Single to Multi-Clouds

[1]M.Bhuvaneshkumar, MCA, Karpagam University [2]G. Anitha, lecturer, Karpagam University

[3]Dr. X. Agnes Kalarani, Professor, Karpagam University

[4]R. Gunasundari, Assistant Professor, Karpagam University

ABSTRACT

The use of cloud computing has increased rapidly in many organizations. Cloud computing provides many benefits in terms of the low cost and accessibility of data. Ensuring the security of cloud computing is a major factor in the cloud computing environment, as users often store sensitive information with cloud storage providers but these providers may be un-trusted. Dealing with "single cloud" providers is predicted to become less popular with customers due to risks of service availability failure and the possibility of malicious insiders in the single cloud. A movement towards "multi-clouds", or in other words, "inter clouds" or "cloud-of-clouds" has emerged recently. This paper surveys recent research related to single and multi-cloud security and addresses possible solutions and also it proves that the maintenance of security in multi-cloud is efficient and easy when compared to single cloud. This work aims to promote the use of multi-clouds due to its ability to reduce security risks that affect the cloud computing user.

INTRODUCTION

Cloud Computing has emerged as a very well-known technique to support large and voluminous data with the help of a shared pool of resources and large storage area. "Cloud computing is a new computing paradigm that is built on virtualization, distributed computing, utility computing and service-oriented architecture. Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

The increasing maturity of cloud computing technology is leading many organizations to migrate their IT infrastructure and/or adapting their IT solutions to operate completely or partially in the cloud. Even governments and companies that maintain critical infrastructures (e.g., health care, telcos) are adopting cloud computing as a way of reducing costs. Nevertheless, cloud computing has limitations related to security and privacy, which should be accounted for, especially in the context of critical applications. Cloud computing, indeed, is a wide-ranging term that transmits hosted services over the Internet. These hosted services are generally separated into three broad categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). The internet is usually represented as the "Cloud".

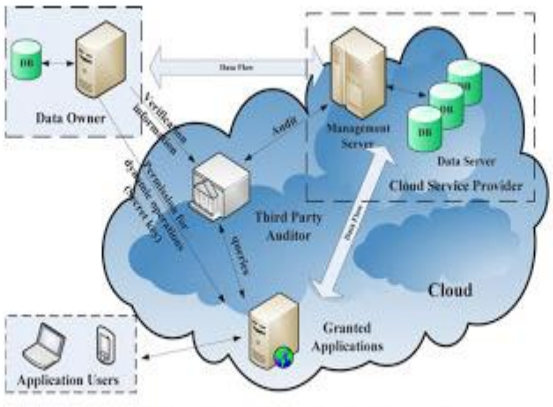
Security in cloud services is based on the following:

- Strong network security is possible around the service delivery platform
- Data encryption: for data in transit (particularly over wide area networks), and sometimes stored data, but it cannot be applied to data in use.
- Access controls to ensure that only authorized users gain access to applications, data and the processing environment and is the primary means of securing cloud-based services

EXISTING SYSTEM:

Cloud providers should address privacy and security issues as a matter of high and urgent priority. Dealing with "single cloud" providers is becoming less popular with customers due to potential problems such as service availability failure and the possibility that there are malicious insiders in the single cloud. In recent years, there has been a move towards "multi-clouds", "inter-cloud" or "cloud-of-clouds".

SYSTEM ARCHITECTURE:



Audit system architecture for cloud computing.

DISADVANTAGES OF EXISTING SYSTEM:

1. Cloud providers should address privacy and security issues as a matter of high and urgent priority.

2. Dealing with "single cloud" providers is becoming less popular with customers due to potential problems such as service availability failure and the possibility that there are malicious insiders in the single cloud

ALGORITHM USED:

- **Secret Sharing Algorithms:**

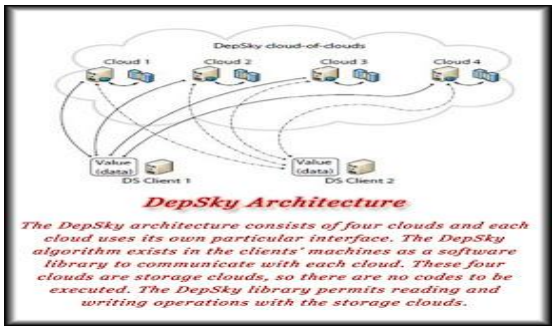
Data stored in the cloud can be compromised or lost. So, we have to come up with a way to secure those files. We can encrypt them before storing them in the cloud, which sorts out the disclosure aspects. Shamir's Secret Sharing Algorithms:

Data stored in the cloud can be compromised or lost. So, we have to come up with a way to secure those files. We can encrypt them before storing them in the cloud, which sorts out the disclosure aspects.However, what if the data is lost due to some catastrophe befalling the cloud service provider? We could store it on more than one cloud service and encrypt it before we send it off. Each of them will have the same file. What if we use an insecure, easily guessable password to protect the 2012 45th Hawaii International Conference on System Sciences file, or the same one to protect all files? I have often thought that secret sharing algorithms could be employed to good effect in these circumstances instead.

PROPOSED SYSTEM:

This paper focuses on the issues related to the data security aspect of cloud computing. As data and information will be shared with a third party, cloud computing users want to avoid an un-trusted cloud provider. Protecting private and important information, such as credit card details or a patient's medical record from attackers or

malicious insiders is of critical importance. In addition, the potential for migration from a single cloud in a multi-cloud environment is examined and research related to security issues in single and multi-clouds in cloud computing are surveyed.



The DEPSKY system model contains three parts: readers, writers, and four cloud storage providers, where readers and writers are the client"s tasks. Bessani et al. explain the difference between readers and writers for cloud storage. Readers can fail arbitrarily (for example, they can fail by crashing, they can fail from time to time and then display any behavior) whereas, writers only fail by crashing

ADVANTAGES OF PROPOSED SYSTEM:

- Data Integrity
- Service Availability.
- The user runs custom applications using the service provider's resources
- Cloud service providers should ensure the security of their customers' data and should be responsible if any security risk affects their customers' service infrastructure.

- **Data Integrity:**

One of the most important issues related to cloud security risks is data integrity. The data stored in the cloud may suffer from damage during transition operations from or to the cloud storage provider. Cachin et al. gives examples of the risk of attacks from both inside and outside the cloud provider, such as the recently attacked Red Hat Linux's distribution servers.

One of the solutions that they propose is to use a Byzantine fault-tolerant replication protocol within the cloud. Hendricks et al. State that this solution can avoid data corruption caused by some components in the cloud. However, Cachinet al. Claim that using the Byzantine fault tolerant replication protocol within the cloud is unsuitable due to the fact that the servers belonging to cloud providers use the same system installations and are physically located in the same place.

It is not an easy task to securely maintain all essential data where it has the need in many applications for clients in cloud computing. To maintain our data in cloud computing, it may not be fully trustworthy because client doesn't have copy of all stored data. But any authors don't tell us data integrity through its user. So we have to establish new proposed system for this using our data reading protocol algorithm to check the integrity of data before and after the data insertion in cloud. Here the security of data before and after is checked by client with the help of CSP using our "effective automatic data reading protocol from user as well as cloud level into the cloud" with truthfulness.

- **Data Intrusion:**

According to Garfinkel, another security risk that may occur with a cloud provider, such as the Amazon cloud service, is a hacked password or data intrusion. If someone gains access to an Amazon account password, they will be able to access all of the account's instances and resources. Thus the stolen password allows the hacker to erase

all the information inside any virtual machine instance for the stolen user account, modify it, or even disable its services. Furthermore, there is a possibility for the user's email(Amazon user name) to be hacked (see for a discussion of the potential risks of email), and since Amazon allows a lost password to be reset by email, the hacker may still be able to log in to the account after receiving the new reset password.

The importance of data intrusion detection systems in a cloud computing environment. We find out how intrusion detection is performed on Software as a Service, Platform as a Service and Infrastructure as Service offerings, along with the available host, network and hypervisor-based intrusion detection options. Attacks on systems and data are a reality in the world we live in. Detecting and responding to those attacks has become the norm and is considered due diligence when it comes to security

- **Service Availability:**

Another major concern in cloud services is service availability. Amazon is mentioned in its licensing agreement that it is possible that the service might be unavailable from time to time. The user's web service may terminate for any reason at any time if any user's files break the cloud storage policy. In addition, if any, damage occurs to any Amazon web service and the service fails, in this case there will be no charge to the Amazon Company for this failure. Companies seeking to protect services from such failure need measures such as backups or use of multiple providers.

DEPSKY, a dependable and secure storage system that leverages the benefits of cloud computing by using a combination of diverse commercial clouds to build a cloud-of-clouds. In other words, DEPSKY is a virtual storage cloud, which is accessed by its users by invoking operations in several individual clouds.

Loss of availability: Temporary partial unavailability of the Internet is a well-known phenomenon. When data is moved from inside of the company network to an external data center, it is inevitable that service unavailability will be experienced. The same problem can be caused by denial-of- service attacks, like the one that allegedly affected a service hosted on Amazon EC2. DEPSKY deals with this problem by exploiting replication and diversity to store the data on several clouds, thus allowing access to the data as long as a subset of them is reachable.

Loss and corruption of data: There are several cases of cloud services losing or corrupting customer data. For example, in October 2009 a subsidiary of Microsoft, Danger Inc., lost the contacts, notes, photos, etc., of a large number of users of the Sidekick service. The data was recovered several days later, but the users of Magnolia were not so lucky in February of the same year, when the company lost half a terabyte of data that it never managed to recover. DEPSKY deals with this problem using Byzantine fault-tolerant replication to store data on several cloud services, allowing data to be retrieved correctly even if some of the clouds corrupt or lose data.

Loss of privacy: The cloud provider has access to both the data stored in the cloud and meta data like access patterns. The provider may be trustworthy, but malicious insiders are a wide-spread security problem. This is an especial concern in applications that involve keeping private data like health records. An obvious solution is the customer encrypting the data before storing it, but if the data is accessed by distributing applications this involves running protocols for key

distribution (processes in different machines need access to the cryptographic keys). DEPSKY employs a secret sharing scheme and erasure codes to avoid storing clear data in the clouds and to improve the storage efficiency

CONCLUSION:

It is clear that although the use of cloud computing has rapidly increased, cloud computing security is still considering the major issue in the cloud computing environment. Customers do not want to lose their private information as a result of malicious insiders in the cloud. In addition, the loss of service availability has caused many problems for a large number of customers recently. Furthermore, data intrusion leads to many problems for the users of cloud computing. The purpose of this work is to survey the recent research on single clouds and multi- clouds to address the security risks and solutions. We have found that much research has been done to ensure the security of the single cloud and cloud storages.

REFERENCES:

1. Mohammed A. Alzain, Eric Parded, BenSoh, James A. Thom, “Cloud Computing Security: From Single to Multi-Clouds”. 2012, ieee conference on system sciences

2. I. Abraham, G. Chockler, I. Keidar and D. Malkhi, "Byzantine disk paxos: optimal resilience with Byzantine shared memory", Distributed Computing, 18(5), 2006, pp. 387-408.

3. G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson and D. Song, "Provable data possession at untrusted stores", Proc. 14th ACMConf. on Computer and communications security, 2007, pp. 598 -609.

4. H. Abu-Libdeh L. Princehouse and H. Weatherspoon "RACS: a case for cloud storage diversity", SoCC'10:Proc. 1st ACM symposium onCloud computing, 2010, pp. 229-240

5. Byzantine disk Pax os: optimal resilience with Byzantine shared memory. Distributed Computing, 18(5):387– 408, April 2006.

6. Hussam Abu-Libdeh, Lonnie Princehouse, and Hakim Weatherspoon. RACS: A case for cloud storage diversity. Proc. of the 1st ACM Symposium onCloud Computing, pages 229–240, June 2010.

7. Hagit Attiya and Amir Bar-Or. Sharing memory with semi-Byzantine clientsand faulty storage servers. In Proc. of the 22rd IEEE Symposium on ReliableDistributed Systems - SRDS 2003, pages 174–183, October 2003.

8. Alysson N. Bessani, Eduardo P. Alchieri, Miguel Correia, and Joni S.Fraga. DepSpace: a Byzantine fault-tolerant coordination service. In Proc. of the 3rd ACM European Systems Conference – EuroSys’08, pages 163–176, April 2008.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

EFFECTIVE DATA MINING USING NEURAL NETWORKS

[1]A.Ishwarya [2]Mrs. P.Laura Juliet M.C.A.,M.Phi

[1]Research Scholarb Vellalar College for Women(Autonomous)
[2]Assistant Professor Vellalar College for Women(Autonomous)
Thindal,Erode.

ABSTRACT

Classification is one of the data mining problems receiving great attention recently in the database community. This paper presents an approach to discover symbolic classification rules using neural networks. Neural networks have not been thought suited for data mining because how the classifications were made is not explicitly stated as symbolic rules that are suitable for verification or interpretation by humans. With the proposed approach, concise symbolic rules with high accuracy can be extracted from a neural network. The network is first trained to achieve the required accuracy rate. Redundant connections of the network are then removed by a network pruning algorithm. The activation values of the hidden units in the network are analyzed, and classification rules are generated using the result of this analysis. The effectiveness of the proposed approach is clearly demonstrated by the experimental results on a set of standard data mining test problems.

1. INTRODUCTION

One of the data mining problems is *classification.* Various classification algorithms have been designed to tackle the problem by researchers in different fields such as mathematical programming, machine learning, and statistics. Recently, there is a surge of data mining research in the database community. The classification problem is re-examined in the context of large databases. Unlike researchers in other fields, database researchers pay more attention to the issues related to the volume of data. They are also concerned with the effective use of the available database techniques, such as efficient data retrieval mechanisms. With such concerns, most algorithms proposed are basically based on decision trees. The general impression is that the neural networks are not well suited for data mining. The major criticisms include the
following:

1) Neural networks learn the classification rules by many passes over the training data set so that the learning time of a neural network is usually long.

2) A neural network is usually a layered graph with the output of one node feeding into one or many other nodes in the next layer. The classification process is buried in both the structure of the graph and the weights assigned to the links between the nodes. Articulating the classification rules becomes a difficult problem.

3) For the same reason, available domain knowledge is rather difficult to be incorporated to a neural network.

On the other hand, the use of neural networks in classification is not uncommon in machine learning community. In some cases, neural networks give a lower classification error rate than the decision trees but require longer learning time. In this paper, we present our results from applying neural networks to mine classification rules for large databases with the focus on articulating the classification rules represented by neural networks. The contributions of our study include the following:

(1) Different from previous research work that excludes the neural network based approaches entirely, we argue that those approaches should have their position in data mining because of its merits such as low classification error rates and robustness to noise.

(2) With our rule extraction algorithms, symbolic classification rules can be extracted from a neural network. The rules usually have a comparable classification error rate to those generated by the decision tree based methods. For a data set with a strong relationship among attributes, the rules extracted are generally more concise.

(3) A data mining system based on neural networks is developed. The system successfully solves a number of classification problems in the literature.

Our neural network based data mining approach consists of three major phases:

(1) Network Construction and Training: This phase constructs and trains a three layer neural network based on the number of attributes and number of classes and chosen input coding method.

(2) Network Pruning: The pruning phase aims at removing redundant links and units without increasing the classification error rate of the network. A small number of units and links left in the network after pruning enable us to extract concise and comprehensible rules.

(3)Rule Extraction: This phase extracts the classification rules from the pruned network. The rules generated are in the form of "if *(a1, Bv1,) and* (x2, *Bv2,) and ... and (xn, Bvn,)* then *Cy* where *a,s* are the attributes of an input tuple, *v* ,a~re constants, & are relational operators (=, <, *2,* <>), and *Ci* is one of the class labels.

2. EXTRACTING RULES FROM A TRAINED NEURAL NETWORK

Network pruning results in a relatively simple network. However, even with a simple network, it is still difficult to find the explicit relationship between the input tuples and the output tuples. A number of reasons contribute to the difficulty of extracting rules from a pruned network. First, even with a pruned network, the links may be still too many to express the relationship between an input tuple and its class label in the form of *if* . . . *then* ... rules. If a network still has *n* input links with binary values, there could be as many as 2, distinct input patterns. The rules could be quite lengthy or complex even for a small *n*. Second, the activation values of a hidden unit could be anywhere in the range 1-1, 11 depending on the input tuple. It is difficult to derive an explicit relationship between the continuous activation values of the hidden units and the output values of a unit in the output layer

2.1 A Rule Extraction Algorithm

The rule extraction algorithm, RX, consists of the four steps given below,

Rule Extraction **algorithm (RX):**

1) Apply a clustering algorithm to find clusters of hidden node activation values.

2) Enumerate the discredited activation values and compute the network outputs Generate rules that describe the network outputs in terms of the discredited hidden unit activation values.

3) For each hidden unit, enumerate the input values that lead to them and generate a set of rules to describe the hidden units' discredited values in terms of the inputs.

4) Merge the two sets of rules obtained in the previous two steps to obtain rules that relate the inputs and outputs.

The first step of **RX** clusters the activation values of hidden units into a manageable number of discrete values without sacrificing the classification accuracy of the network. After clustering, we obtain a set of activation values at each hidden node. The second step is to relate these discredited activation values with the output layer activation values, i.e., the class labels. And the third step is to relate them with the attribute values at the nodes connected to the hidden node. A general purpose algorithm X2R was developed and implemented to automate the rule generation process. It takes as input a set of discrete patterns with the class labels and produces the rules describing the relationship between the patterns and their class labels.

To cluster the activation values, we used a simple clustering algorithm which consists of the following steps,

1) Find the smallest integer d such that if all the network activation values are rounded to d-decimal-place, the network still retains its accuracy rate.

2) Represent each activation value a by the integer nearest to a*10 pow(4).LetH={H1,H2,….Hn} be the set of these discrete representations. Set i=1.

3) Sort the set H such that the values of Hi are in increasing order.

4) Find a pair of distinct adjacent values hi,j and hi,j+l in Hi such that if hi,j+l is replaced by hi,j, no conflicting data will be generated.

5) If such values exist, replace hi,j+1 by hi,j and repeat step 4. Otherwise set i=i+1, if i<=H go to step 3.

This algorithm tries to merge as many as possible values into one interval in every step as long as the merge do conflicting data, i.e., two sets of activation values that are the same but they represent tuples that belong to different classes. The algorithm is order sensitive, that is, the resulting clusters depend on the order in which the hidden units are clustered.

3. CONCLUSION

In this paper we present a neural network based approach to mining classification rules from given databases. The approach consists of three phases:

(1) Constructing and training a network to correctly classify tuples in the given training data set to required accuracy

(2) Pruning the network while maintaining the classification accuracy; and

(3) Extracting symbolic rules from the pruned network.

The work reported here is our attempt to apply the connectionist approach to data mining to generate rules similar to that of decision trees. A number of related issues are to be further studied. One of the issues is to reduce the training time of neural networks. Although we have been improving the speed of network training by developing fast algorithms, the time required to

extract rules by our neural network approach is still longer than the time needed by the decision tree based approach. As the long initial training time of a network may be tolerable in some cases, it is desirable to have incremental training and rule extraction during the life time of an application database. With an incremental training that requires less time, the accuracy of
rules extracted can be improved along with the change of database contents. Another possibility to reduce the training time and improve the classification accuracy is to reduce the number of input units of the networks by feature selection.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, ”Database Mining: A Performance Perspective,” *IEEE Trans. Knowledge and Data Eng.,* vol. 5, no. ***6,*** Dec. 1993.

[2] J.E. Dennis Jr. and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Englewood Cliffs, N.J.: Prentice Hall, 1983.

[3] H. Lu, R. Setiono, and H. Liu, “Neurorule: **A** Connectionist Approach to Data Mining,” *Proc, VLDB* ’95, pp. 478-489,1995.

[4] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, *Machine Learning, Neural and Statistical Classification.* Ellis Horwood Series in Artificial Intelligence, 1994.

[5] H. Liu and S.T. Tan, ”X2R A Fast Rule Generator,” *Pvoc. IEEE Int’l Conf.* Systems, *Man, and Cybernetics.* IEEE, 1995.

[6] J.R. Quinlan, *C4.5: Programs **for** Machine Learning.* Morgan Kaufmann, 1993.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

MULTIMODAL BIOMETRIC SCHEME USING FINGERPRINT & IRIS FUSION

V. Sridevi [1], P.Vidhya [2]

[1]Assistant Professor, Department of Computer Applications
Dr. N.G.P.Arts and Science College, Coimbatore-48

[2]Research Scholar, Department of Computer Science,
Dr. N.G.P. Arts and Science College, Coimbatore-48.

Abstract

In certain applications based on multimodal interaction it may be crucial to determine not only what the user is doing, but who is doing it, in order to prevent fraudulent use of the system. The biometric technology, and particularly the multimodal biometric systems, represents a highly efficient automatic recognition solution for this type of applications. Although multimodal biometric systems have been traditionally regarded as more secure than unmoral systems, their vulnerabilities to spoofing attacks have been recently shown. New fusion techniques have been proposed and their performance thoroughly analyzed in an attempt to increase the robustness of multimodal systems to these spoofing attacks. However, the vulnerabilities of multimodal approaches to software-based attacks still remain unexplored. In this work we present the first software attack against multimodal biometric systems. Its performance is tested against a multimodal system based on face and iris, showing the vulnerabilities of the system to this new type of threat. Score quantization is afterwards studied as a possible countermeasure, managing to cancel the effects of the proposed attacking methodology under certain scenarios Gabor filters responses are widely and successfully used as general purpose features in many computer vision tasks, such as in texture segmentation, face detection and recognition, and iris recognition. In a typical feature construction the Gabor filters are utilized via multi-resolution structure, consisting of filters tuned to several different frequencies and orientations. Hence, single biometric may not be able to achieve the desired performance requirement in real world applications. One of the methods to overcome these problems is to make use of multimodal biometric authentication systems, which combine information from multiple modalities to arrive at a decision.

Keywords: Multimodal Biometric – Gabor Filter – Fingerprint Recognition – Face Recognition

I. *INTRODUCTION*

Data mining has been used intensively and extensively by many organizations. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Data mining applications can greatly benefit all parties involved in the healthcare industry.



(a) Enhanced image. (b) Binirization image. (c) Thinned image.

A biometric system provides automatic identification of an individual based on a unique feature or characteristic possessed by the individual. Automated measurement of biological or/and behavioral characteristics of person for medical, security or psychological purposes. Depending on the application context, a biometric system may be called either a verification system or an identification system. A verification system verifies a person by comparing the captured biometric characteristic with his own biometric template pre-stored in the system. It conducts one-to-one comparison to determine whether the identity claimed by the dividable is true. A verification system either rejects or accepts the submitted claim of identity [22].

Fingerprints have been an accepted method in forensic investigations to identify suspects and repeat criminals. Now, virtually all law enforcement agencies worldwide use



Automatic Fingerprint Identification Systems (AFIS). With growing concerns about terrorist activities, security breaches, and financial fraud, other physiological and behavioral human characteristics have been used for person identification.

However, it is not easy to extract the minutiae points accurately from the original fingerprint images and the performance of the feature extraction algorithm relies heavily on the quality of the input images. It is essential to implement fingerprint enhancement process before extracting minutiae for the robustness of fingerprint identification algorithm with respect to the quality of fingerprint images. Fingerprint enhancement is intended to reduce noises and improve the contrast between ridges and valleys in the ray-scale fingerprint images [23].

II. *RELATED WORKS*

The general term face recognition can refer to different application scenarios. One scenario is called recognition or identification, and another is called authentication or verification. In either scenario, face images of known persons are initially enrolled into the system. This set of persons is sometimes referred to as the gallery. Later images of these or other persons are used as probes to match against images in the gallery.

In a recognition scenario, the matching is one-to-many, in the sense that a probe is matched against the entire gallery to find the best match above some threshold. In an authentication scenario, the matching is one-to-one, in the sense that the probe is matched against the gallery entry for a claimed identity, and the claimed identity is taken to be authenticated if the quality of match exceeds some threshold. The recognition scenario is more technically challenging than the authentication scenario. One reason is that in a recognition scenario a larger gallery tends to present more chances for incorrect recognition. Another reason is that the whole gallery must be searched in some manner on each recognition attempt.

The 3D shape of the face is often sensed in combination with a 2D intensity image. In this case, the 2D image can be thought of as a 'texture map' overlaid on the 3D shape. An example of a 2D intensity image and the corresponding 3D shape are shown in Fig. 2.1, with the 3D shape rendered in the form of a range image, a shaded 3D model and a mesh of points. A range image, also sometimes called a depth image, is an image in which the pixel value reflects the distance from the sensor to the imaged surface.

III. *PROBLEM DEFINITION*

The increasing advancements in the field of Information Technology & World Wide Web, causing human beings frequently confront with various kinds of unauthorized access. Also with the enlargement of mankind's activity range, the importance for person's status identity is becoming more and more important. So many different techniques for person's status identity have been proposed for this practical task.

Conventional methods for status identity check like password and identification card are not always reliable, because these methods can be easily forgotten, stolen or forged. A wide variety of biometrics have been developed for this challenge, examples include automatic retinal vasculature scan, iris recognition, fingerprints matching, hand shape identification, handwritten signature verification, and voice recognition systems.

Iris recognition is considered to be the most reliable and accurate biometric identification system available. Iris recognition system captures an image of an individual's eye, the iris in the image is then meant for the further Segmentation and normalization for extracting its Feature. The performance of iris recognition systems depends on the process of segmentation. Segmentation is used for the localization of the correct iris region in the particular portion of an eye and it should be done accurately and correctly to remove the eyelids, eyelashes, reflection and pupil noises present in iris region.

Score quantization is afterwards studied as a possible countermeasure, managing to cancel the effects of the proposed attacking methodology under certain scenarios Gabor filter responses are widely and successfully used as general purpose features in many computer vision tasks, such as in texture segmentation, face detection and recognition, and iris recognition. In a typical feature construction the Gabor filters are utilized via multi-resolution structure, consisting of filters tuned to several different frequencies and orientations.

IV. *MAJOR CONTRIBUTION*

The following evaluation techniques are the main contributions of the work, which has been screened out in this paper:

- Both the global flow pattern of ridges and valleys and local characteristics (inter-ridge distances, ridge orientation) are used for feature extraction.
- They generate a short fixed length code, Finger Code (Feature vector) for fingerprint Finger Code is suitable for fast matching (by Euclidean distance), storage on smartcard and indexing the obtained representation is scale, translation and rotation invariant.

A. Methodology

In the identification stage, each input image is associated with a registered user of the system. In the verification stage one must decide if it really represents a face of that person. It propose an automatic threshold-based verification approach Thus, it compute a proper threshold value and compare the enrage distances from each face class with it. If the average distance corresponding to an image from a class is greater than threshold T, then that image is invalidated and rejected from the face class. The verification process is represented formally as

$$\forall i \in [1, N], \quad \forall I \in C_i : \frac{\sum_{j=1}^{n(i)} d(V(I), V(F_j^i))}{n(i)} > T \Rightarrow C_i = C_i - \{I\} \quad (1)$$

1. Binarization

Most minutiae extraction algorithms operate on binary images where there are only two levels of interest: the black pixels that represent ridges, and the white pixels that represent valleys. Binarization is the process that converts a grey level image into a binary image. This improves the contrast between the ridges and valleys in a fingerprint image, and consequently facilitates the extraction of minutiae. One useful property of the Gabor filter is that it has a DC component of zero, which means the resulting filtered image has a mean pixel value of zero.

2. Thinning

The final image enhancement step typically performed prior to minutiae extraction is thinning. Thinning a morphological operation that successively erodes away the foreground pixels until one pixel wide. A standard thinning algorithm is employed, which performs the thinning operation using two sub Iterations. This algorithm is accessible in MATLAB via the “thin” operation under the bimorph function. Each sub-iteration begins by examining the neighborhood of each pixel in the binary image, and based on particular set of pixel-deletion criteria, it checks whether the pixel can be deleted or not. These sub iterations continue until no more pixels can be deleted.

3. Minutiae Extraction

The most commonly employed method of minutiae extraction is the Crossing Number (CN) concept this method involves the use of the skeleton image where the ridge flow pattern is eight-connected. The minutiae are extracted by scanning the local neighborhood of each ridge pixel in the image using a3x3 window. The CN value is defined as half the sum of the differences between pairs of adjacent pixels in the eight-neighborhood.

$$CN = 0.5 \sum_{i=1}^{8} \left| P_i - P_{i+1} \right| \quad P_9 = P_1$$

Where Pi is the pixel value in the neighborhood of P. For a pixel P, its eight neighboring pixels are scanned in an anti-clockwise direction.

4. Post Processing

False minutiae may be introduced into the image due to factors such as noisy images, and image artifacts created by the thinning process. Hence, after the minutiae are extracted, it is necessary to employ a post processing stage in order to validate the minutiae.

Validation algorithm by Tico and Kuosmanen has chosen to implement. This algorithm tests the validity of each minutiae point by scanning the skeleton image and examining the local neighborhood around the point. The first step in the algorithm is to create an image M of size

WxW, where M corresponds to the WxW neighborhood centered on the candidate minutiae point in the skeleton image. The central pixel of M corresponds to the minutiae point in the skeleton image, and so this pixel is labeled with a value of -1. The rest of the pixels in M are initialized to values of zero.

5. Minutia Match

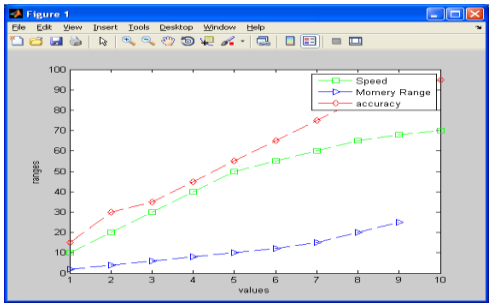
Given two sets of minutia of two fingerprint images, the minutia match algorithm determines whether the two minutia sets are from the same finger or not. An alignment-based match algorithm is used in our system. It includes two consecutive stages: one is alignment stage and the second is match stage.

In the alignment stage, given two fingerprint images to be matched, choose any one minutia from each image; calculate the similarity of the two ridges associated with the two referenced minutia points. If the similarity is larger than a threshold, transform each set of minutia to a new coordination system whose origin is at the referenced point and whose x-axis is coincident with the direction of the referenced point.

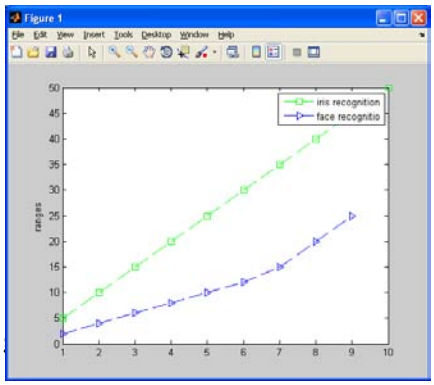
V. *EXPERIMENTAL RESULTS*

The proposed face and iris recognition techniques as well as the other techniques was applied to the test set using all of the random realizations of each training set. The speed, memory range and accuracy criteria are also examined efficiently using graph. In the fig.1, the speed range from 15 and proceeds to the range above 80 (i.e) the speed of recognizing the face and iris is high. Then the memory are also predicted, it ranges from 10. It will gradually increases depend upon the biometric input.

Fig.1 Speed and accuracy, memory ranges in biometric recognition,



Fi



.

2 Recognition Range of Face & Iris

Then the accuracy also explored to screen out the efficiency of biometric recognition. The iris and face recognition range estimated are shown in fig.2.

VI. *CONCLUSION*

A high recognition rate has been achieved by our technique, as it results from the performed experiments. The obtained results prove the effectiveness of our method. This technique provides a higher recognition rate than many other facial recognition approaches. We have compared it with our Eigen image based recognition technique, and found they produce similar results for identical face datasets. Our future work will continue the research in the facial recognition domain. Since the described face recognition approach provides a satisfactory person authentication, we intend to include it into a more complex biometric system that uses more identifiers besides human face. Also, we will focus our research on face detection and face localization domains .Thus we can combine this face recognition technique with a face detection method, to obtain a system that is able to recognize faces from image scenes and video sequences.

In order to improve the automatic segmentation algorithm, a more elaborate eyelid and eyelash detection system could be implemented. An improvement could also be made in the speed of the system. The most computation intensive stages include performing the Hough transform, and calculating Hamming distance values between templates to search for a match. Since the system is implemented in MATLAB,

which is an interpreted language, speed benefits could be made by implementing computationally intensive parts in C or C++.

REFERENCES

1. W. Zhao, R. Chellappa, P. J. Phillips, "Face Recognition: A Literature Survey", ACM
2. M. A. Turk, A. Pentland, "Face Recognition Using Eigenfaces", In Proc. Of Computer Vision And Pattern Recognition, IEEE, Pp. 586–591, 1991.
3. T. Barbu, "Eigenimage-Based Face Recognition Approach Using Gradient Covariance", Numerical Functional Analysis.
4. X. Y. Jing, H. S. Wong, D. Zhang, "Face Recognition Based On 2d Fisher Face Approach", Pattern Recognition, 39, 4, Pp. 707–710, April 2006.
5. F. Samaria, S. Young, "HMM Based Architecture for Face Identification", Image and Computer Vision, 12, Pp. 537–583, October 1994.
6. L. Wiskott, C. Malsburg, "Face Recognition By Dynamic Link Matching", In J. Sirosh, R. Miikkulainen And Y. Choe Editors, "Lateral Interactions In The Cortex: Structure And Function", UTCS Neural Networks Research Group, Austin, TX, 1996,Electronic Book, ISBN 0-9647060-8.
7. C. Liu, H. Wechsler, "Gabor Feature Classifier For Face Recognition", In Proceedings of the ICCV, Volume 2, Pp. 270-275, 2001.
8. J. R. Movellan, Tutorial on Gabor Filterst. Acharya, A. K. Ray, "Image Processing – Principles And Applications", Wiley Interscience, 2005
9. Variable Lighting", IEEE Trans. Pattern Anal. Mach. Intelligence, 27, 5, Pp. 684–698, 2005.
10. M. H. Yang, D. Kriegman, N. Ahuja, "Detecting Faces in Images: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence (Pami), 24, 1, Pp. 34–58, Jan. 2002.
11. W. Zhao, R. Chellappa, P. J. Phillips, "Face Recognition: A Literature Survey", Acm Computing Surveys, 35, 4, Pp. 399–458,December 2003.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Color Image Segmentation for Fruit Ripeness Detection

Jothipriya.I
Assistant Professor of Computer Science
S.R.N.M.College of Arts and Science

ABSTRACT

In this paper, we are representing different techniques to detect the rate of ripeness of fruits and vegetables. This paper reports techniques like histogram matching, clustering algorithms based image segmentation and relative value of parameter based segmentation. Each technique uses coloured images of fruits and vegetables as input data. In these techniques we set some threshold levels. By comparing the input data image with these threshold levels we can find the maturity level of given fruits and vegetables.

Keywords-- Image Segmentation, Ripeness, Clustering

I. INTRODUCTION

IN older days, human depends upon its vision qualities to differentiate between ripe and unripe fruits. But this method had high rate of errors because of illness, distraction and other factors during working hours [6]. This also may effects the working speed of system. So to decrease this failure rate human started to invent new methods.

These days, there are various methods to detect the ripeness of fruits and vegetables. In some methods we apply chemicals on fruits and sometimes we use machines. As we know, chemical may effects human health so usually machines are used for this purpose. Machine use their visual-based colour classification system that provide reliability, high speed and repeatable operation. Hence the production increases and reduces its dependency on manpower.

In machine vision system computer uses different method to analyse the given image of fruit and vegetable. Previously, computer systems were not robust enough to operate on large and real colours of images, so mostly gray scale images had been the main focus for researchers. But today, computer system has been developed enough to work on large and true colour images [2].

To increase the efficiency of computer system different researchers perform various experiments on different fruits and vegetables to check their maturity levels.

B. Ojeda-Magana, et al used different partitional clustering algorithms to detect ripeness level of bananas and tomatoes [1]. Fatma Susilawati mohamad and Azizah Abdul Manaf used histogram matching method to find ripeness of oil palm fruits. Scanlon proposed an approach to quantify colour of potato chips [3] and Choi used Colour image analysis to detect tomatoes maturity rate [8]. In 2011, Chiunhsiun Lin, et al proposed a method to check the maturity rate of tomatoes [2].

This paper is organized as follows; section 2 gives description about colour image segmentation. Section 3 depicts the outline of some techniques used to find the maturity levels of fruits and vegetables. Finally, some conclusions are represented in 4.

II. COLOUR IMAGE SEGMENTATION

Colour image segmentation is a process of partitioning an image into meaningful regions with respect to colours. Firstly monochromatic images are used to perform image segmentation operation. But in these images intensity is the only information source. It has been said that human eye can recognize thousands of colour shades and intensities but in case of gray scale images it can recognize only two dozens of gray shades. So segmentation of colour images has been more preferred as compare to gray scale images. It is easy to segment image on basis of colours as compare to texture, shape and size.

The main reason behind this colour images provide more information, more capacity and high speed to process the information [1].

With the development of technology, the research work in field of colour image segmentation also increases. So according to techniques applied, this research work can be categorized as: 1. Image domain based, 2. Physics based, 3.Feature space based [9].

Image domain based techniques are mainly used to provide image segmentation for closely connected regions [9]. In first type of approach we try to develop regions until an homogeneity constrain is held. This development of region can be done with neighbouring pixels or by merging and splitting regions. In second type of approach we detect image discontinuities, so regions are limited by the pixels containing the colour boundary. The resulting segmentation may or may not be homogeneous region in feature space.

Physics based techniques use models based on the physical properties of light and of objects in an image. Such techniques attempt to find actual material boundaries and ignore boundaries due to illumination changes in an image [10].

In feature space based techniques algorithms are concerned with the presence of collective massive pixels within a features spectral space [9]. After determined relevant colour classes, the segmentation have been done according to regions of pixels corresponding to same colour class. But the final segmentation result is not so visible because of some limitations. These methods don't use its neighbourhood pixel information.

III. DIFFERENT TECHNIQUES TO DETECT RIPENESS OF FRUITS

There are various techniques to detect the ripeness of fruits and vegetables. Three of them are following:

1) Clustering algorithms

Clustering means to create groups of same elements whereas algorithm means a step by step procedure for calculation. Hence clustering algorithms are the step by step procedure to calculate clusters of similar colour pattern in image. In [1], there are many partitional clustering algorithms that are used for colour image segmentation like K- means, the Fuzzy c means(FCM), the Gustafson Kessel improved by Babuska(GK-B) and the Gustafson Kessel Possobilistic Fuzzy c Means(GKPFCM). The performance of these algorithms depends upon the amount or type of information and the distance measure it uses. In this case, distance is the squared or absolute difference between a pixel and cluster centre but the result can be improved if spatial information is also taken into account. This will improve the better identification and quantification of the objects in the partitioned region of image.

We can check the results of these algorithms after applying them on fig 1(a). This is the original image of ripe and unripe bananas. Fig 1(b) shows the result of K-means algorithm. Fog 1(c) shows the result of Fuzzy c-means algorithm.

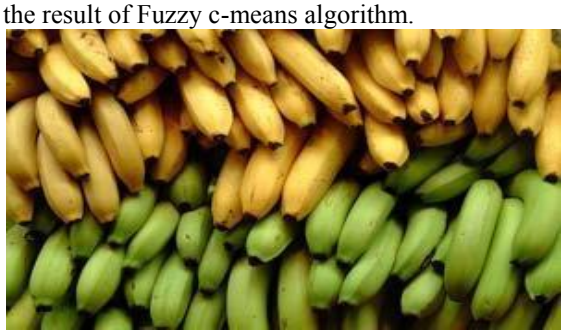


Fig 1(a) Original image



Fig 1(b)

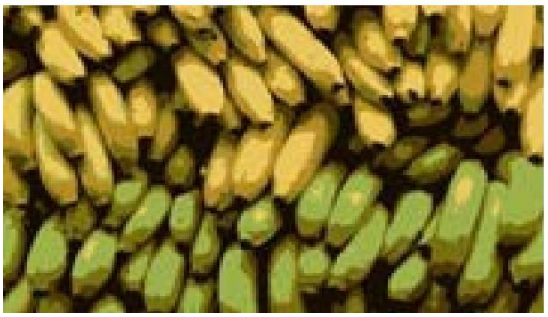


Fig 1(c)



Fig 1(d)



Fig 1(e)

Fig 1(d) shows the result of GK-B (Gustafson-kessel-babuska) algorithm and finally fig 1(e) shows the result of GKPFCM (Gustafson-kessel Possibilistic Fuzzy c-Means) algorithm. At first sight we can see that the last algorithm shows the best result.

2) Histogram matching

In [14], according to author "Colour histogram represent the distribution or division of colours in an image". Global contrast of many images can be increases with this method, especially when the usable data of the image is represented by close contrast values. These adjustments help in uniformly distribution of intensities on the histogram. Hence the areas of lower local contrast can gain higher contrast without affecting the global contrast [12].

In this method we convert the coloured image of fruits into gray scale image and find its histogram. After that we compare that histogram with the histogram of ripe fruit histogram sample and by this comparison we can easily find the ripeness of given fruit. According to [11], the colour histogram term is more often used for three dimensional spaces like RGB or HSV. But colour histogram use for all colour spaces. In [13] different histogram similarity measures are used to find most suitable and efficient one to check the ripeness of oil palm fruits.

3) Parameter based segmentation

According to [2], in this technique RGB colour space is used and the whole processing is directly applied on original coloured images. Firstly we set the value of three parameters - red, green and blue colours to detect ripe fruits. For example we know the ripe tomato have red colour and unripe tomato have green colour. To detect the ripeness of tomato in given image we will set the values of RGB colours in such a manner that only red colour will be segmented. Consider r (i), g (i) and b (i)) represent the intensities of red, green, and blue light. So the three basic rules for colour image segmentation of tomatoes are:

1. $r(i)>\alpha$, means primary colour component red should be greater than α.
2. $\beta1<(r(i)-g(i))<\beta2$, means the primary colour component (red-green) should be between $\beta1$ and $\beta2$.
3. $\gamma1<(r(i)-b(i))<\gamma2$, means the primary colour component (red-blue) should be between $\gamma1$ and $\gamma2$.

The first rule means that the value of r (i) — the intensity of red light should be larger than α. The second rule means that the value of (r (i) – g (i)) — (the intensity of red light) - (the intensity of green light) should be between $\beta1$ and $\beta2$. The third rule means that the value of (r (i) – b (i)) — (the intensity of red light) - (the intensity of blue light) should be between $\gamma1$ and $\gamma2$.

So, to calculate the maturity level of tomatoes in [2] author set the values as: $\alpha=20$, $\beta1=0$, $\beta2=110$, $\gamma1=20$, $\gamma2=160$. At last we can calculate the rate of maturity level of tomato by formula R = (the number of maturity tomatoes colour) / (the number of NOT the maturity tomatoes colour). More values of R mean more maturity.

In Fig 2(a) the queue of tomatoes are arranged according to their maturity levels i.e. lest most tomato is ripe one and the right most is unripe. After applying this technique on original image 2(a), segmented image 2(b) is obtained.

IV. CONCLUSION

Clustering algorithms performance can be improved if we use more quantity of information. It also depends upon type of distance measure used. Histogram matching technique provides better result because it increases the global contrast of image so that intensities can be better distributed on histogram. But both clustering algorithms and histogram matching needs colour space transform whereas parameter based segmentation don't use any colour space transform. So the last technique is also very effective to find maturity levels of fruits and vegetables without difficulty by partially changing the values of parameters (α, $\beta1$, $\beta2$, $\gamma1$, $\gamma2$).

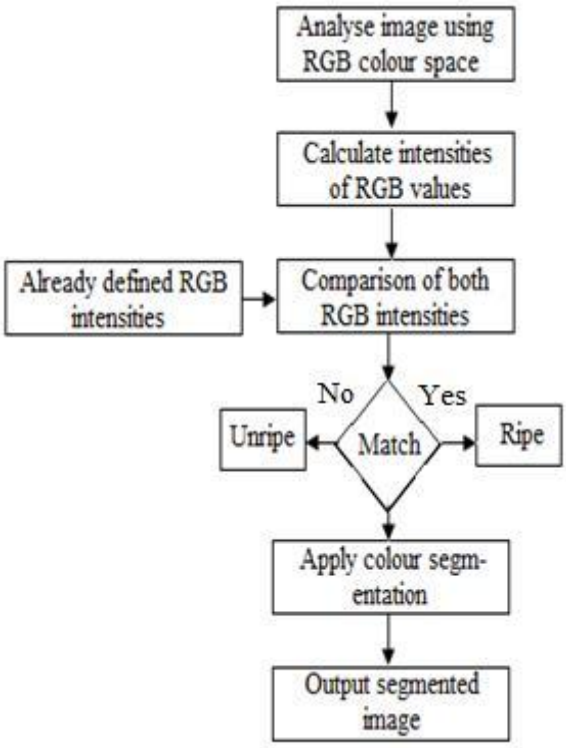


Fig. 2 Block diagram of Parameter based

Image segmentation



Fig 2(a)



Fig 2(b)

REFERENCES

[1] B. Ojeda-Magana, R. Ruelas, J. Quintanilla-Domingvez and D. Andina, *"Colour Imge Segmentation by Partitional Clustering Algorithns"*, IECON 2010 - 36th Annual Conference on IEEE .Industrial Electronics Society, 7-10 Nov., Glendale, AZ, pp.2828 - 2833, 2010.

[2] Chiunheium Lin, ching- Hung Su,Hsuan Shu Huang and Kuo-Chin Fan, *"Colour Image Segmentation Using Relative values of RGB in Various Illumination, Circumstances"*, International Journal of computers, vol.5(2),pp.252-261, 2011.

[3]Scanlon, M. G., "*Computerized video image analysis to quantify colour of potato chips"*, American Potato Journal, Vol. 71(11), pp. 717-733, 1994.

[4]F. Mendoza, J .M. Aguilera, "*Application of Image Analysis for Classification of Ripening Bananas",* Journal of Food Science. 69(9): pp.E471-E477, 2004.

[5]J. Blasco, N. Aleixos, E., "Molto. *Machine Vision System for Automatic Quality Grading of Fruit",* Biosystems Engineering, vol. 85 (4):pp.415-423, 2003.

[6]Ling Mei Chang, Rodney Tan, Gilbert Thio, "*Design Of Visual-based Colour Classification System"*, Research papers, JASA 2,pp. 30-33, January 2007.

[7]Zhi- yuan Wen, Lu-ming Shen, Hui-png jing, Kui Fang, "*Colour and Shape Grading Of Citrus Fruit Based on Machine Vision with Fractel Dimention",* 3rd International Congress on Image and Signal Processing (CISP2010),pp. 898-903,IEEE 2010.

[8] Choi K. H., "*Tomato maturity evaluation using colour image analysis"*, Transactions of the ASAE, Vol. 38(1), pp. 171-176, 1995.

[9] S. Romani, P. Sobrevilla, E. Montseny, "*Obtaining the Relevant Colours of an image through Stability- based Fuzzy Colour Histograms"*, The IEEE International Conference On Fuzzy Systems,pp. 914-919,2003.

[10] Allan Hanbury, "*Physics based Segmentation of Colour Images in Spherical Coordinates",* Technical Report, pp. 1-2,23rd july, 2003.

[11] Wikipedia, the free encyclopedia. Color Histogram, 2010.

[12] "Histogram Equalization", Wikipedia The Free Encyclopedia, 2010.

[13]Fatma Susilawati Mohamad, Azizah Abdul Manaf & Suriayati Chuprat, "*Histogram Matching For Color Detection:A Preliminary Study",*Information Technology (ITSim), 2010 International Symposium in , Kuala Lumpur ,Vol. 3,pp. 1679 – 1684,*„IEEE* 2010

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Subject Based Clustering Of Web-Data Using Association Rule Mining

[1]Ms M.Kalpana, [2]Dr.V.Arulmozhi

[1]Research scholar , Department of Computer Science, Tiruppur Kumaran College For Women, Tirupur 641 687. Bharathiar University,Coimbatore, Tamil Nadu, India.

[2]Associate Professor, Department of Computer Applications, Tiruppur Kumaran College For Women, Tirupur 641 687. Bharathiar University,Coimbatore, Tamil Nadu, India.

Abstract

As the World Wide Web is growing drastically in day-by-day, the processing speed and storage of data needs efficient algorithms for information retrieval. The present search engines accompanied with indexer algorithms to provide adequate results for user query. This paper proposing the method of clustering web data using association rule mining for improving efficiency of search engine.

Key words: *Subject-based clustering, Association rule mining, Web data mining, Search engine algorithm.*

1. Introduction

Clustering is an unsupervised learning approach of machine learning. In data mining it comes under descriptive mining of data. Clustering involves grouping the similar objects in the same class known as cluster. It is one of the main tasks of data mining. Researchers using various clustering algorithms for their research based on area of interest. Popular clustering algorithm is small distance based algorithm. The most widely used applications are machine learning, pattern recognition , Image analysis, Information retrieval and bio-informatics.

In [12] the process of clustering was used to produce homogeneous result based on user query. The clustered search results followed particular characteristics will give faster and better result for the user [12]. The homogenous words which doesn't care about grammar and semantic. For example "connection", "connecting"," connected" will be described by the root word "connect". In [10] the documents in the web were clustered based on text mining which incorporating concept based clustering by using frequency and semantic of the documents.

Association rules involve to uncover interesting and frequent relationships between data in data sets. Since it introduced in 1993 [6] for discovering the relationship in the large database of market analysis by customer transaction.

In [4] the association rule clustering system was useful for segment the data in cluster.

Association rule mining defined as,

Let I = $\{i_1,i_2\ldots i_n\}$ be a set of n attributes called as *items*. Let D = $\{t_1,t_2\ldots t_m\}$ be a set of transactions in database D. Each unique transaction ti in data base D contains a subset of items in I. A rule can be defined as X=> Y where $X,Y \subseteq I$ and $X \cap Y = \emptyset$ which can be defined by IF-THEN rule, if (*ancedent)* X comes then Y (*consequent)* will definitely followed the X. For example in Table 1 when the user search about "Noun" the ancedent, , must also search about "Verb" the consequent.

Association rule uses the interestingness measuring concepts like support and confidence. The support of X=>Y is the percentage of transactions in the data base contains XUY in total transactions. Confidence or strength for an association rule X=>Y is ratio of the number of transactions that contain XUY to the number of transactions that contain X. Other measuring concepts also there such as lift, conviction, correlation , regression etc.

Algorithms for association rules are developed based on data sets [3]. Apriori algorithm is pioneer to all algorithms of association rules

which using support-based pruning in data set. Brute-force is an effective method for generating candidate item sets. In [8] the highly ranked association rules for clustering the data.

2. Applications of Association Rule Mining (ARM)

The association rules can be employed in various areas like market basket analysis, Medical diagnosis, Bio- medical literature, protein sequence, Census data, Logistic regression, Fraud detection in web [2]. ARM (Association Rule Mining) can be used for student assessment data in knowledge discovery [3]. ARM used for pattern discovery [1] in large base for improving website design, advertisement and related link connections.ARM can be adapted for various databases such as large database, distributed database, medical database, relational database and spatial database.

3. Subject-Based Clustering Using Association Rule Mining (ARM).

TABLE 1
SURVEY DATA BASED ON BROWSING TOPICS.

Topic (X)	Related Topics (Y)
Noun	{Verb, Pronoun, Adverb, Adjective}
Algebra	{Variables, Constants, Co-efficient, Solution}
Population Census	{Migration, Birth rate, Death rate}
Electricity	{Magnetism, Fusion, Nuclear Fusion}
Metals	{Iron, Steel, Alloy}
Pollution	{Soil pollution, water pollution, Noise pollution}
Solar system	{ Sun, Planets, Earth, Satellite}

TABLE 2
BINARY REPRESENTATION DATA

T id	Noun	Verb	Pro noun	Adverb	Adjective
1	1	1	1	0	0
2	1	1	0	1	0
3	0	0	1	0	1
4	1	1	0	1	1
5	0	0	0	1	1

TABLE 3
SUPPORT OF ASSOCIATION RULES

T id	X => Y	Support
1	Noun => Verb	60%
2	Noun => Pronoun	20%
3	Noun => Adjective	20%
4	Verb => Adverb	40%
5	Adverb => Adjective	40%

In this paper we proposing the system of clustering the web data based on association rules which involve high intra cluster tasks. The present search engines using indexer algorithms [11] by key words to retrieving data for user query which contain list of related links in first 20-50 results displayed. With the proposed system we can reduce the next retrieval of user query. For example the user who search about “Noun“should search about “Verb” in Table 1.

In [11] the topic map was introduced for clustering the documents on web using three assertions such as name of file, occurrence and associations. Document clustering captures the semantic of the text for retrieve the particular site. Subject based information management [11] on web using clustering outperforms the previous semantic based clustering document.

The above mentioned subject based information management can be enhanced by the subject-based clustering with ARM. The rule can be obtained by user interaction with the search engine [9] to make the community of users. Off line clustering also done by some sort of survey or the association between web pages based on subject. Online clustering can be done by using search queries.

3.1 Sample Data Analysis

Table 1 shows the survey conducted in the group of 20 teachers of various subjects. The survey is about what their searching criteria in particular topics.

Table 2 shows the Binary representation [3] of the topic noun in English subject. Table 3 represents supporting percentage of the transactions in Table1.

The support of X UY where X={Noun} and Y={Verb} calculated by the division of item sets which contain XUY by total number of transactions. Support of XUY is 3/5 = 0.6 which is 60%.

The confidence of X=> Y is calculated by Supp(XUY)/ Supp(X), 0.6/0.6=1 which is equal to 100% so the user search about "Noun" will search "Verb" as well.

The Apriori algorithm using support ≥ minimum support [6] to find the association rule in large data base.

Based on above calculations the search results of the query "Noun" will also contain link for "Verb" so that the user would not move for next query which reduce the one access to the web data base.

4. Conclusion and Future Enhancement

In this paper we discussed about clustering, association rules and their applications. We proposed the concept of clustering the web data using association rule for search engine algorithm which reduce more accesses to the web database.

Future enhancement could be developing efficient algorithm for the proposed system.

Acknowledgement

First author would like to thank Dr.V.Arulmozhi, Assistant professor, Department of Computer Applications ,Tiruppur kumaran college for women, Tirupur, for her invaluable support and the Second author would like to thank Tiruppur Kumaran College for Women, Tiruppur.

REFERENCES

[1]. Ms Kiruthika M, Mr Rahul Jadhav , Ms Rashmi J, Ms Dipa Dixit , Ms Anjali Nehete P, Ms Trupti Khodkar "*Pattern Discovery Using Association Rules "* International journal of advanced computer science and applications, Vol 2 No. 12 , 2011.

[2]. Sotiris Kotsiantis, Dimitris Kanellopoulos," *Association Rules Mining: A Recent Overview",* GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[3]. Dr.Varun Kumar , Anupama Chandha,"*Minining Association Rules in Students Assessment Data",* International journal of Computer Science Issues Vol 9, Issue 5 , No 3, September 2012.

[4]. V.Umarani . Dr.M. Punithavalli ,"*A Study of Effective Mining of Association Rules From Huge Database",* International journal of Computer Science and Research Vol 1, issue 1, 2010.

[5]. Sachin Sharma , Veenu Mangat ," *Current Scenario of Powerful Clustering Technique",* International Association of Scientific Innovation and Research (IASIR) (An Association Unifying the Sciences, Engineering, and Applied Research) ISSN (Online): 2279-0071.

[6]. Rakesh Agarwal , Tomasz .Imie.Linski, Arun Swamy,"*Mining association Rules Between Set Of Items in Large databases".* Proceeding of ACM-SIGMO, International conference, May 1993.

[7]. Dr.M.Renuka devi, Mrs. A. Baby Sarojini, "*Applications of Association Rule Mining in Different databases",*Journal of Global research in Computer Science Vol 3, No 8, August 2012, ISSN 2229-371X.

[8]. S.Kannan, R.Bhaskaran ,"*Association Rule Pruning Based on Interestingness Measures With Clusterings"*,International journal of Computer Science Issues , Vol 6, No 1, 2009, ISSN 1694-0784.

[9]. Jianning Pong, William Perrizo, Qin Ding and Jingkai Zhou, "*The Application of Association Rule To Remotely Sensed Data".* ACM Symposium on Applied Computing, Como, Italy, 2000, pp 340-345.

[10]. V.M.Navaneethakumar, Dr.C.Chandrasekar ,"*A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model"*, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012 ,ISSN (Online): 1694-0814.

[11]. Muhammed Rafi, M.Shahid Shaik,Amir Farroq," *Document Clustering Based on Topic Maps",* International Journal of Computer Applications(0975-8887), vol 12-No 1,December 2010.

[12]. Adina LIPAI,"*World Wide Web Metasearch Clustering Algorithm",*Revista Information Economatica. Nr.2(46)/2008.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Content-based Recommendation System for Interactive Digital Televisions

[1]Mrs. R.Kousalya, Phd scholar, [2]Saranya.K,
[1]Manonmaniam Sundaranar University, Head, Dept of Computer Applications,
[2] Research Scholor ,Department of Computer Science ,
Dr.N.G.P Arts and Science College,Coimbatore-48.

ABSTRACT

Recommender systems arose in view of the information overload present in numerous domains. The so called content-based recommenders offer products similar to those the users liked in the past. We propose a personalized recommender system that captures the user's viewing habits without obstructing the usual way TV is watched by implicit feedback method. Our proposal describes program representation and retrieval, user modeling, and aggregation of her/his estimated interests. We aim that our proposed method promptly learns the user's preferences and delivers valued recommendations.

INTRODUCTION:

Information overload is becoming increasingly severe in our modern times of omnipresent mass-media and global communication facilities, hampering the user's ability to distinguish relevant information from irrelevant. Consequently, during the last years, significant research efforts have been striving to conceive automated filtering systems that provide users with desirable and interesting information [1]. In the last decade, the so-called recommender systems have been gaining momentum as another efficient means of reducing complexity when searching for relevant information. Besides, recommenders have also attracted increasing public interest, leveling the ground for new business opportunities in different fields, such as e-commerce and Digital TV [2, 3]. Their main objective is to sift through very large sets of information and to select those products the users will appreciate, based on their past preferences, history of purchase, or demographic information.

For that purpose, recommender systems which can be running either remotely in a server or locally in a domestic or mobile consumer device employ several personalization strategies. The first such strategy was *content-based filtering*, which suggests to a user products similar to those he/she liked in the past. In spite of its accuracy, this technique is limited due to the employed similarity metrics. These metrics are based on syntactic approaches that only detect resemblance between products that share all or some of their attributes [4].

Consequently, traditional approaches lead to *overspecialized* suggestions including only products that bear a strong resemblance to those the user already knows (i.e. products with attributes that are already defined in his/her profile). The existing recommender systems fight *overspecialization* by a strategy named collaborative filtering, based on offering to each user products that have been appealing to others with similar preferences (named his/her neighbours). Collaborative filtering alleviates *overspecialization*, but it also causes new limitations, such as scalability problems, difficulties to select each user's neighbourhood when the available preferences are sparse, and privacy concerns related to the confidentiality of the users' personal data (see [5] for details).

We observe TV broadcast for the following reasons: it is a widely spread communications service, it has well-established usage patterns, and it is currently facing an expansion of available content. The restriction to this particular service does not affect the generality of our analysis or applicability of the results. To help their users find the programs of interest, many cable operators nowadays offer electronic program guides (EPGs) as a part of standard subscription packages. These usually provide nothing more than a list of programs that are aired at a given time; in that sense, they do not differ from print TV magazines. What the users would truly benefit from is a personalized program guide, which would

track and learn their viewing habits – without causing any noticeable discomfort – and then recommend them only those programs they were likely to be interested in. This is a typical task for recommender systems [6].

We identify three properties of the existing TV broadcast which influence the design and operation of personalized program guides, i.e. lack of return channel, limited user interface and well-established usage habits. The lack of return (uplink) channel means that the user equipment (TV receiver or set-top-box) cannot send the acquired data about user's viewing preferences to the network center, which further means that it is not possible to consider the viewing patterns of other users. It is argued that TV recommender systems should therefore rely on the techniques that could be applied locally in the user's TV receiver [7,8]. This eliminates the otherwise superior collaborative and hybrid recommenders and leaves the content-based techniques, known from information retrieval, as methods of choice for personalized EPGs.

TECHNOLOGICAL LANDSCAPE:

In order to fulfill the goals of our strategy it is necessary to combine technologies from diverse areas of information and communications technology:

• *Semantic annotation of TV programs*. Firstly, the TV programs available during the recommendation process must be annotated by metadata that describe accurately their main attributes.

• *Formalization of domain knowledge*. The knowledge provided by the annotations of the TV programs must be formalized to carry out the semantic reasoning processes that lead to our enhanced recommendations. For that purpose, we exploit the mechanisms developed in the Semantic Web, an initiative in which information is given a well-defined meaning, better enabling computers and people to work in cooperation [9].

• *TV platforms*. Finally, it is also necessary to bear in mind the characteristics of the TV platforms for the domestic receivers and mobile devices used to access the suggested TV programs, considering the technical aspects of the available broadcast networks.

CONTENT-BASED STRATEGY:

With the goal of diversifying the recommendations in the TV domain, our personalization strategy suggests programs that are semantically associated with the contents the viewer has liked in the past, disregarding the syntactic similarity metrics [10] adopted in the traditional content-based methods. Specifically, our reasoning-based strategy consists of two stages – named *filtering phase* and *recommendation phase*, respectively– which are sketched below and described next:

• **Filtering phase**: Broadly speaking, our content-based strategy selects the instances of classes and properties that are relevant for the user, by considering his/her personal preferences. This module exploits the user profile to suggest relevant items by matching the profile representation against that of items to be recommended. The result is a binary or continuous relevance judgment (computed using some similarity metrics [11]), the latter case resulting in a ranked list of potentially interesting items.

• **Recommendation phase**: The knowledge inferred from the filtering phase is processed in the second phase. This intelligent mechanism is able to: (i) explore efficiently the entities and associations discovered during the filtering phase, and (ii) select among them the TV programs suggested to the user.

Our approach models the user's profiles based on the user's preferences and interests. Specifically, we propose a semantic model for each user that provides information about:

(i) the TV programs that were appealing and uninteresting for him/her (named *positive* and *negative preferences*[12], respectively),
(ii) their main attributes, and
(iii) the genres under which these programs are classified.

This user modeling approach has two main strengths:

On the one hand, our profiles provide a formal representation of the users' preferences, being able to reason about them and discover extra knowledge about their interests. Such knowledge permits to compare effectively the users' preferences with the available products, thus leading to personalization processes more accurate than the traditional syntactic approaches [5]. On the other hand, our approach does not require defining in each user's profile the classes, properties and instances that identify his/her preferences [13], thus reducing the storage capabilities required in the recommender system.

The concept applied to content-based filtering try to find the most relevant documents based on the user's behavior in the past. Such approach however restricts the user to documents similar to those already seen. This is known as the over-specialization problem. The interests of a user are rarely static but change over time. Instead of adapting to the user's interests after the system has received feedback one could try to predict a user's interests in the future and recommend documents that contain information that is entirely new to the user.

To overcome the over-specialization problem and to enhance the user's viewing experience, three adaptive feedback schemes were proposed in [14]. The experimental results showed that the implicitly positive feedback scheme achieved the best performances, both in terms of success rate and user comfort.

RESULT AND DISCUSSION:

In the paper, we presented a framework for recommender systems in TV which includes user profile modeling, aggregation of user's interests and program retrieval. It is estimated that the proposed method is efficient when used in interactive TV platforms. It is also discussed that the proposed work, when implemented with implicit feedback method would be more effective with simpler interaction with the user.

CONCLUSION:

We have proposed a method of Content-based recommender systems for TV. The personalization strategy overcomes unresolved limitations of the existing TV program recommender systems. Moreover, we also observed that our method can provide more suitable programs or user preferences, even when the number of recommended results is small. Further, we aim at resolving the over-specialization problem by using the implicit-feedback method.

REFERENCES:

[1] K. van Rijsbergen, "Information retrieval". University of Glasgow, 1975.
[2] B. Schafer, J. Konstan, and J. Riedl, "Recommender systems in ecommerce," Proc. of 1st ACM Conference on Electronic Commerce, 1999, pp. 158- 166.
[3] A. Griffiths, "Digital television strategies: business challenges and opportunities". Palgrave Macmillan, 2003.
[4] M. Montaner, B. López, and J. de la Rosa, "A taxonomy of recommender agents on the Internet," Artificial Intelligence Review, vol. 19, no. 4, pp. 285-330, June 2003.
[5] G. Adomavicius, and A. Tuzhilin, "Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," IEEE Trans. on Knowledge and Data Engineering, vol. 17, no. 6, pp. 739-749, Jan. 2005.
[6] TV-Anytime Specification Series S-3 on Metadata, 2001 [on line]. Available in http://www.tv-anytime.org.
[7] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," The Scientific American, vol. 279, no. 5, pp. 34-43, May 2001.
[8] S. Staab, and R. Studer, Handbook on Ontologies (International Handbook on Information Systems). Springer, 2003.
[9] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6, pp. 734-749, June 2005.
[10] L. Ardissono et al., "User modeling and recommendation techniques for personalized electronic program guides," in: L. Ardissono et al. (eds.) Personalized Digital Television, Kluwer Academic Publishers, 2004, pp. 3-26.
[11] T. Isobe et al., "Development of a TV reception system personalized with viewing habits", IEEE Trans. Consumer Electron., vol. 51, no. 2, pp. 665-674, May 2005.
[12] A. L. Buczak, J. Zimmerman, K. Kurapati, "Personalization: Improving ease-of-use, trust and accuracy of a TV show recommender," Proc. TV02 Workshop on Personalization in TV, Malaga, 2002, pp. 3-12.
[13] J. Barneveld, M. Setten, "Designing usable interfaces for TV recommender systems," in L. Ardissono et al. (eds.) Personalized Digital Television, Kluwer Academic Publishers, 2004, pp. 259-285.
[14] Milan Bjelica , Ana Perić , "Adaptive Feedback Schemes for Personalized Content Retrieval" IEEE Transactions on Consumer Electronics, Vol. 57, No. 3, August 2011



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Unprivileged Black-Box Detection of User-Space Key loggers

[1]P.K.Mangayarkarasi MCA.,MPhil., Assistant Professor [2]S.Narmadha, Research scholar,

[1 & 2] Kongu Arts and Science College, Erode.

ABSTRACT

Software keyloggers are a fast growing class of invasive software often used to harvest confidential information. One of the main reasons for this rapid growth is the possibility for unprivileged programs running in user space to eavesdrop and record all the keystrokes typed by the users of a system. The ability to run in unprivileged mode facilitates their implementation and distribution, but, at the same time, allows one to understand and model their behavior in detail. Leveraging this characteristic, we propose a new detection technique that simulates carefully crafted keystroke sequences in input and observes the behavior of the keylogger in output to unambiguously identify it among all the running processes. We have prototyped our technique as an unprivileged application, hence matching the same ease of deployment of a keylogger executing in unprivileged mode. We have successfully evaluated the underlying technique against the most common free keyloggers. This confirms the viability of our approach in practical scenarios. We have also devised potential evasion techniques that may be adopted to circumvent our approach and proposed a heuristic to strengthen the effectiveness of our solution against more elaborated attacks. Extensive experimental results confirm that our technique is robust to both false positives and false negatives in realistic settings.

Index Terms—Invasive software, keylogger, security, black-box, PCC

INTRODUCTION

KEYLOGGERS are implanted on a machine to intentionally monitor the user activity by logging keystrokes and eventually delivering them to a third party [1]. While they are seldom used for legitimate purposes (e.g., surveillance/parental monitoring infrastructures), keyloggers are often maliciously exploited by attackers to steal confidential information. Many credit card numbers and passwords have been stolen using keyloggers [2], [3], which makes them one of the most dangerous types of spyware known to date.Keyloggers can be implemented as tiny hardware devices or more conveniently in software. Software-based keyloggers can be further classified based on the privileges they require to execute. Keyloggers implemented by a kernel module run with full privileges in kernel space. Conversely, a fully unprivileged keylogger can be implemented by a simple user-space process. It is important to notice that a user-space keylogger can easily rely on documented sets of unprivileged APIs commonly available on modern operating systems (OSs). This is not the case for a keylogger implemented as a kernel module. In kernel space, the programmer must rely on kernel-level facilities to intercept all the messages dispatched by the keyboard driver, undoubtedly requiring a considerable effort and knowledge for an effective and bug-free implementation. Furthermore, a keylogger implemented as a user-space process is much easier to deploy since no special permission is required.

MODERN KEYLOGGERS

Breaching the privacy of an individual by logging his keystrokes can be perpetrated at many different levels. For example, an attacker with physical access to the machine might wiretap the hardware of the keyboard. A dishonest owner of an Internet cafe´, in turn, may find it more convenient to purchase a software solution, install it on all the terminals, and have the logs dropped on his own machine. Depending on the setting, a keylogger can be implemented in many different ways. For instance, external keyloggers rely on some physical property, either the acoustic emanations produced by the user typing [6], or the electromagnetic emanations of a wireless keyboard [7]. Hardware keyloggers are still external devices, but are implemented as dongles placed in between keyboard and motherboard. All these strategies, however, require physical
access to the target machine. To overcome this limitation, software approaches are more commonly used. Hypervisor-based keyloggers

are the straightforward software evolution of hardwarebased keyloggers, literally performing a man-in-the-middle attack between the hardware and the operating system.

Kernel keyloggers come second in the chain and are often implemented as part of more complex rootkits. In contrast to hypervisor-based approaches, hooks are directly used to intercept buffer-processing events or other kernel messages.

User-space keyloggers, on the other hand, do not require any special privilege to be deployed. They can be installed and executed regardless of the privileges granted. User-space keyloggers can be further classified based on the scope of the hooked message/data structures. Since a system hosts multiple applications, keystrokes can be intercepted either globally (i.e., for all the applications) or locally (i.e., within the application).Weterm these two classes of user-space keyloggers type I and type II. Fig. 1 shows the proposed classification: the left pane shows the process of delivering a keystroke to the intended application.

User-space keyloggers can be further classified based on the scope of the hooked message/data structures. Since a system hosts multiple applications, keystrokes can be intercepted either globally (i.e., for all the applications) or locally (i.e., within the application).Weterm these two classes of user-space keyloggers type I and type II.

INJECTOR

The role of the injector is to inject the input stream into the system, simulating the behavior of a user at the keyboard. By design, the injector must satisfy several requirements.
First, it should only rely on unprivileged API calls. Second, it should be capable of injecting keystrokes at variable rates to match the distribution of the input stream. Finally, the resulting series of keystroke events produced should be no different than those generated by a real user. In other words, no user-space keylogger should be somehow able to distinguish the two types of events. To address all these issues, we leverage the same technique employed in automated testing.

MONITOR

The monitor is responsible to record the output stream of all the running processes. As done for the injector, we allow only unprivileged API calls. In addition, we favor strategies to perform realtime monitoring with minimal overhead and the best level of resolution possible. Finally, we are interested in application-level statistics of I/O activities, to avoid dealing with filesystem-level caching or other potential nuisances. Fortunately, most modern operating systems provide unprivileged API calls to access performance counters on a per-process basis.

DETECTOR

The success of our detection algorithm lies in the ability to infer a cause-effect relationship between the keystroke stream injected in the system and the I/O behavior of a keylogger process, or, more specifically, between the respective patterns in AKP form. While one must examine every candidate process in the system, the detection algorithm operates on a single process at a time, identifying whether there is a strong similarity between the input pattern and the output pattern obtained from the analysis of the I/O behavior of the target process. Specifically, given a predefined input pattern and an output pattern of a particular process, the goal of the detection algorithm is to determine whether there is a match in the patterns and the target process can be identified as a keylogger with good probability.

ALGORITHMS USED

Random (RND). Each sample is generated at random.

Random with fixed range (RFR). The pattern is a random permutation of a series of samples uniformly distributed .This is to maximize the amount of variability in the input pattern.

Impulse (IMP). Every sample 2i is assigned the value of 0 and every sample 2i þ 1 is assigned the value of 1. This algorithm attempts to produce an input pattern with maximum variance and idle periods at minimum.

Sine wave (SIN). The pattern generated is a discrete sine wave distribution oscillating between 0 and 1. The sine wave grows or drops with a fixed step of 0.1. This algorithm explores the effect of constant increments and decrements in the input pattern.

KEYLOGGER DETECTION

To evaluate the ability to detect real-world keyloggers, we experimented with all the keyloggers from the top monitoring free software list an online repository continuously updated with reviews and latest developments in the area.

All the keyloggers were detected within a few seconds without generating any false positives; in particular, no legitimate process scored PCC values _ 0:3. Virtuoza Free Keylogger required a longer window of observation to be detected; this sample was indeed the only keylogger to store keystrokes in memory and flush out to disk at regular time intervals. Nevertheless, we were still able to collect consistent samples from flush events and report high PCC values.

In a few other cases, keystrokes were kept in memory but flushed out to disk as soon as the keylogger detected a change of focus. This was the case for Actual Keylogger, Revealer Keylogger Free, and Refog Keylogger Free. To deal with this common strategy, our detection system enforces a change of focus every time a sample is injected. In addition, some of the keyloggers examined included support for encryption and most of them used variable-length encoding to store special keys. As Section 5.2 demonstrates, our approach deal with these nuisances transparently with no effect on the resulting PCC. Another potential issue arises from keyloggers dumping a fixed-format header on the disk every time a change of focus is detected. The header typically contains the date and the name of the target application.

CONCLUSION

In this paper, we presented an unprivileged black-box approach for accurate detection of the most common key loggers, i.e., user-space key loggers. We modeled the behavior of a key logger by surgically correlating the input (i.e., the keystrokes) with the output (i.e., the I/O patterns produced by the key logger). In addition, we augmented our model with the ability to artificially inject carefully crafted keystroke patterns. We then discussed the problem of choosing the best input pattern to improve our detection rate. Subsequently, we presented an implementation of our detection technique on Windows, arguably the most vulnerable OS to the threat of key loggers. To establish an OS-independent architecture, we also gave implementation details for other operating systems. We successfully evaluated our prototype system against the most common free key loggers with no false positives and no false negatives reported. Other experimental results with a homegrown key logger demonstrated the effectiveness of our technique in the general case. While attacks to our detection technique are possible and have been discussed at length. We believe our approach considerably raises the bar for protecting the user against the threat of key loggers.

REFERENCES

[1] T. Holz, M. Engelberth, and F. Freiling, "Learning More About the Underground Economy: A Case-Study of Keyloggers and Dropzones," Proc. 14th European Symp. Research in Computer Security, pp. 1-18, 2009.

[2] San Jose Mercury News, "Kinkois Spyware Case Highlights Risk of Public Internet Terminals," http://www.siliconvalley.com/ mld/siliconvalley/news/6359407.htm, 2012.

[3] N. Strahija, "Student Charged After College Computers Hacked," http://www.xatrix.org/article2641.html, 2012.

[4] N. Grebennikov, "Keyloggers: How They Work and How to Detect Them," http://www.viruslist.com/en/analysis? pubid=204791931, 2012.

[5] Security Technology Ltd., "Testing and Reviews of Keyloggers, Monitoring Products and Spyware," http://www.keylogger. org, 2012.

[6] L. Zhuang, F. Zhou, and J.D. Tygar, "Keyboard Acoustic Emanations Revisited," ACM Trans. Information and System Security, vol. 13, no. 1, pp. 1-26, 2009.

[7] M. Vuagnoux and S. Pasini, "Compromising Electromagnetic Emanations of Wired and Wireless.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Secured Networking using Steganography

[1]N. Hemalatha [2]S.Sivaraja M.Sc.,M.Phil.,

[1]Research Scholar, Department of Computer Science

[2] Assistant professor, Kongu arts and science college, Erode.

ABSTRACT

Nowadays the world is a global market. Every business and enterprise is related to the internet and technology in one way or the other way.Networking plays a major role in today's computer technology. But the fullest advantages are exploited only in a secured connection that is Network Security. This paper deals with improvising the security of network connection to prevent unauthorized access using "Steganography" as an interpreting tool.

Introduction:

The world is a global market today. Every business and enterprise is related to the internet and technology in one way or the other way and thus the demand for Network Security has increased many folds. Recent cases of hacking and data theft has left many organizations with no option but to hire the services of experts who can perform specific operations to check how unsafe the data is and how prone the network is to threat from data thieves. To prevent our valuable data so we need good data security techniques called as **Steganography**.

Steganography is the art and science of hiding information by embedding messages within other, seemingly harmless messages. More commonly, steganography is used to supplement encryption. An encrypted file may still hide information using steganography, so even if the encrypted file is deciphered, the hidden message is not seen.

What is Computer Network?

A computer network is a system in which computers are connected to share information and resources. The connection can be done as peer-to-peer or client/server. This web site reviews the techniques you can use to set up and possibly manage a network for home or a small business.



Fig. 1 Example of computer Network

Computer Network security:

Computer security is a branch of computer technology known as Information Security as applied to computers and networks.



The objective of computer security includes protection of information and property from theft, corruption, or natural disaster, while allowing the information and property to remain accessible and productive to its intended users.

Why do we need security?



Fig. 2 Type of Network attacks

Because of these,

In general, the majority of network communications occur in an unsecured or "clear text" format, which allows an attacker who has gained access to data paths in your network to "listen in" or interpret (read) the traffic.

When an attacker is eavesdropping on your communications, it is referred to as sniffing or snooping. The ability of an eavesdropper to monitor the network is generally the biggest security problem that administrators face in an enterprise. Internet Security, among different types of computer security, covers malware and hacking techniques.

The Internet is an open zone where anyone can create a website that may place malware on your computer or server.

Common security attacks and their counter measures:

- ❖ DoS- Denial of Service
- ❖ Trojan Horse - Comes with other software.
- ❖ Virus - Reproduces itself by attaching to other executable files.
- ❖ Worm - Self-reproducing program. Creates copies of itself. Worms that spread using e-mail address books are often called viruses.
- ❖ Logic Bomb - Dormant until an event triggers it (Date, user action, random trigger, etc.).

Hacker Attacks:

The term "hacker attacks" to indicate hacker attacks that are not automated by programs such as viruses, worms, or Trojan horse programs. There are various forms that exploit weakness in security. Many of these may cause loss of service or system crashes.

IP spoofing

Gaining access through source routing



Fig. 3 Example of Hacker

- ✓ Man in the middle attack
- ✓ Server spoofing
- ✓ Password cracking

DoS Attacks:

1).Ping broadcast

Ping request packet is sent to a broadcast network address where there are many hosts.

2) Smurf

An attack where a ping request is sent to a broadcast network address with the sending address spoofed so many ping replies will come back to the victim and overload the ability of the victim to process the replies.

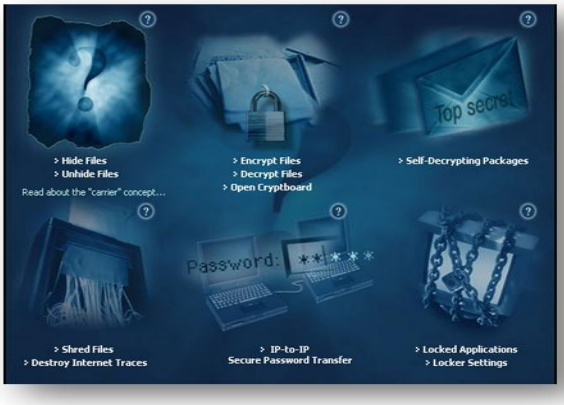


Fig. 4 Type of Network Security

Covert Channel - Stegnography:

Steganography is one of the best known methods of covert communication.

Steganography is the art and science of hiding information by embedding messages within other, seemingly harmless messages. More commonly, steganography is used to supplement encryption. An encrypted file may still hide information using steganography, so even if the encrypted file is deciphered, the hidden message is not seen.

A Brief History of Steganography:

Steganography, or "hidden writing" can be traced back to 440 BC in ancient Greece. Often they would write a message on a wooden panel, cover it in wax, and then write a message on the wax. These wax tablets were already used as writing utensils, so the hiding of a message in a commonly used device draws very little suspicion. In addition to use by the Greeks, the practice of steganography was utilized by spies in World War II.

Role of steganography in Network security:

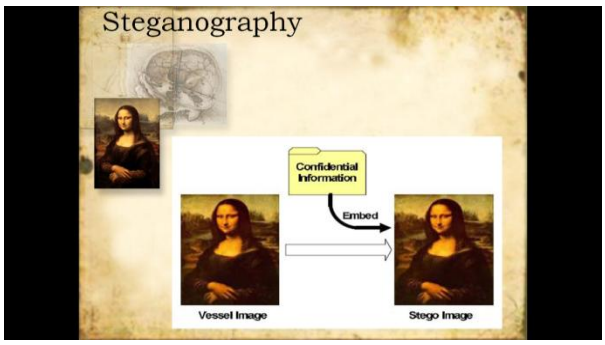


Fig. 6 Example of steganography technic in image

What is it used for?

- ✓ Hiding the fact that you are sending messages
- ✓ Hiding several messages inside data
- ✓ Digital watermarking

Kerckhoffs' principle

- ✓ Secure with knowledge of the system
- ✓ Message can only be read with secret key

Implementation:

Hiding information

- ✓ Text / WebPages
- ✓ Images
- ✓ Audio
- ✓ Video

Text / WebPages

- ✓ Use of a codebook
- ✓ Layout of texts
- ✓ Every Nth character
- ✓ Use of whitespaces and newlines
- ✓ Can be difficult to detect and decode

Ex:

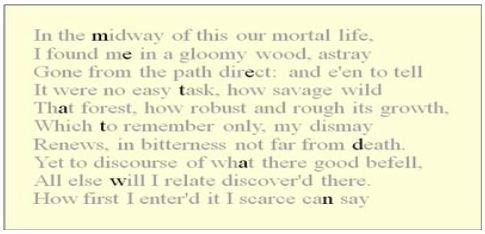

In the midway of this our mortal life,
I found me in a gloomy wood, astray
Gone from the path direct: and e'en to tell
It were no easy task, how savage wild
That forest, how robust and rough its growth,
Which to remember only, my dismay
Renews, in bitterness not far from death.
Yet to discourse of what there good befell,
All else will I relate discover'd there.
How first I enter'd it I scarce can say

Fig. 7 Example of steganography technic in Text

Images

- ✓ JPEG uses discrete cosine transformations (DCT)
- ✓ Transform pixel blocks to DCT coefficients
- ✓ Get least significant bit of each DCT coefficient
- ✓ Replace LSB with secret message bit
- ✓ Insert modified DCT into output image.

Audio

- ✓ Use frequencies inaudible to humans
- ✓ Embed data using the LSB
- ✓ Encoding musical tones

Video

- ✓ Visible signals (gestures, movements)

- ✓ Same techniques as with images
- ✓ Same techniques as with audio

Types of Steganography techniques:

1. Physical steganography

Hidden messages within wax tablets — in ancient Greece, people wrote messages on the wood, and then covered it with wax upon which an innocent covering message was written.

Hidden messages on messenger's body — also used in ancient Greece. Herodotus tells the story of a message tattooed on a slave's shaved head, hidden by the growth of his hair, and exposed by shaving his head again.

During World War II, the French Resistance sent some messages written on the backs of couriers using invisible ink.

2. Digital steganography

The applications of digital steganography in various e-commerce applications through the Internet will be
discussed in detail. These applications include digital watermarking for copyright protection of multimedia data, digital signature authentication and validation of electronic documents, digital data storage and linkage for binding digitized photographs with personal attribute information, as well as secure communication of multimedia data.



Fig.8 Example of Digital steganography

3. Network steganography:

All information hiding techniques that may be used to exchange steganograms in telecommunication networks can be classified under the general term of network steganography.

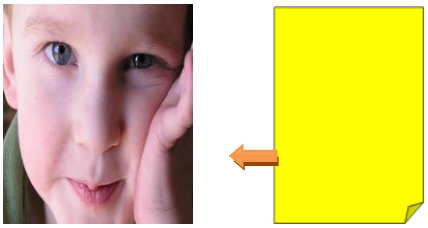


Fig. 9 Example of Network steganography

Single Carrier File
400k in size

Single Carrier File
400k in size

Typical network steganography methods involve modification of the properties of a single network protocol. Such modification can be applied to the PDU.

A simplified example with an 8-bit image

1 pixel:

(00 01 10 11)

White red green blue

Insert 0011:

(00 00 11 11)

White white blue blue

As can be seen from the example, with an 8-bit image, the cover image must be carefully selected since LSB manipulation is not as forgiving because of the color limitations.

A simplified example with a 24-bit image **1 pixel:** (00100111 11101001 11001000) **Insert 101:**

(00100111 11101000 11001001)

red green blue

4. Text steganography

Steganography can be applied to different types of media including text, audio, image and video etc. However, text steganography is considered to be the most difficult kind of steganography due to lack of redundancy in text as compared to image or audio but still has smaller memory occupation and simpler communication.

Benefits of Stegnography:

- ✓ High data security
- ✓ Data integrity
- ✓ Authentication of message originator

- ✓ non-reputation

Conclusion:

- Covert Channels- Stegnography provides the means for communicating without being noticed. They allow you to bypass normal network security mechanisms You can only combat covert channels if you understand how it works,steganographic tools involved hiding messages inside digital images or sound files, known as carriers, like that *Thriller* MP3.

- Now steganography has entered a new era, with stupendously greater potential for mischief. With the latest technique.

- References

[1]Schneier, Bruce (3 March 2006). "Caller ID Spoofing". schneier.com. Retrieved 16 January 2011.

[2]Gantz, John; Rochester, Jack B. (2005). *Pirates of the Digital Millennium*. Upper Saddle River, NJ 07458: Prentice Hall. ISBN 0-13-146315-2.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

DATA MINING - A NOVEL TACTIC

[1]Ms.K.Suguna,[2]Ms.S.Deepa Ranjini,[3].Ms.K.Kalaivani

[1,2,3] Assistant professor
[1,2,3] Department of Computer Applications
Dr.N.G.P Arts and Science College

ABSTRACT

Web mining is used to discover the information from the World Wide Web and their usage patterns. Web usage mining is the one of the type of web mining which allows for the collection of Web access information for Web pages. The binary data is transformed into real data by using wiener transformation algorithm. Using k-means algorithm the transformed data are clustered based on the similarities. Then the Apriori algorithm is used for data preprocessing to find the frequent patterns. In this paper we proposed the FP-Growth Algorithm on web log files to extract the most frequent pattern.

Keywords - World Wide Web, Web Mining, Web Usage Mining, Apriori, Weiner Transformation algorithm, FP Growth.

1 INTRODUCTION

The World Wide Web enabled the spread of information over the Internet through an easy-to-use and flexible format. The Web is a system of interlinked hypertext documents accessed through the Internet. The web mining is used to extract the useful information from the World Wide Web by using data mining techniques. Web mining can be classified into three categories: (i) web content mining, (ii) web structure mining and (iii) web usage mining. The web content mining is used to search the web pages via content. The web structure mining is used discover the web page structure and hierarchy of hyper links in the web site.

The web usage mining is the task of applying data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of users navigating on the Web. It is used to discover the navigation patterns from web data, predicts the user behavior while the user interacts with the web and also it helps to improve large collection of resources.

Web usage mining consists of three main steps: (i) preprocessing, (ii) pattern discovery and (iii) pattern analysis. The data extracted from the web log files are preprocessed to remove the noisy data. The data collection should be done before the preprocessing phase.

In data preprocessing phase the data are cleaned and the frequent patterns are mined. The preprocessed file consists of information such as who accessed the web site, what pages where accessed and how long the user accessed that page. In pattern discovery phase the activities of the users on the web are discovered. The frequent patterns discovery phase needs only the Web pages visited by a given user. In this stage the sequences of the pages are irrelevant. Also the duplicates of the same pages are omitted, and the pages are ordered in a predefined order. In pattern analysis phase the patterns extracted from the pattern discovery phase are processed to get most frequent pattern.

2 WEB USAGE MINING

Web usage mining is the process of finding the navigation patterns and their use of web resources. The data collection, data preprocessing, pattern discovery and pattern analysis are the major task of the web usage mining. [2] new

2.1 Data Collection

The data collection phase contains the raw data from the data source. In this paper the binary database is the data source. The binary database should be transformed into real data by using wiener transformation algorithm to make the data to ready for the pre processing.

2.2. Data preprocessing

a) Data cleaning: The web log file may consist of some irrelevant data so it is necessary to remove that unwanted data. In this phase that unwanted data are removed from the given dataset.

b) Path completion: In this path completion phase the user access paths are identified and the missing paths are added.

c) Session Identification: The timeout method is used to identify the individual user session, if the page request time exceeds a certain limit that implies user is started a new session.

d) User Identification: Users are identified, who contact web server, requesting for some resource on the web.

2.3. Apriori

Apriori Algorithm is an influential algorithm for mining frequent item sets for boolean association rules. The Apriori algorithm is used in data mining process for mining frequent patterns from the given data set. This algorithm uses an iterative approach called level-wise search, in which n-item sets are used to explore n+1 item sets. The set of frequent 1_itemsets, frequent 2_itemsets and frequent 3_itemsets are found until no more frequent n_itemsets can be found.

Some of the issues of Apriori algorithm are: Database scanning of the whole dataset for each iteration, the computational efficiency is very less because the whole database scans is needed every time, the cost of generating large number of candidate sets and scanning the database repeatedly. The repeated scan of the database is very costly.

To overcome these issues a new method Weiner transformation algorithm is introduced.

a) Apriori: In Apriori algorithm, the binary data is transformed into real domain using linear wiener transformation, based on its neighborhood property. The Wiener transformed data is partitioned into K clusters using the multi-pass K-means algorithm. Apriori procedure is used for the K similar groups of data from which, frequent item sets can be generated and association rules are derived. Large datasets are partitioned so that the candidate item sets generated will be very less and database scanning will also be done for adequate data which increases the efficiency. [1]The weiner transformation algorithm is described in Algorithm1.

Algorithm 1

Function wiener2 (X_i)

Input : Binary data vector X_i of size 1 X q

Output : Transformed data vector Y_i of size 1 X q

Step 1: Calculate the mean μ for the input vector X_i around each element

$$\mu = \frac{1}{pq} \sum_{n_1, n_2 \in \eta} X(n_1, n_2)$$

Where η is the local neighborhood of each element

Step 2: Calculate the variance σ^2 around each element for the vector

$$\sigma^2 = \frac{1}{pq} \sum_{n_1, n_2 \in \eta} X^2((n_1, n_2) - \mu)$$

Where η is the local neighborhood of each element

Step 3: Perform wiener transformation for each element in the vector using equation Y based on its neighborhood

$$Y(n_1, n_2) = \mu + \frac{\sigma^2 - \lambda^2}{\sigma^2} (X(n_1, n_2) - \mu)$$

Where λ^2 is the average of all the local estimated variances.

CLUSTER 1:

Support 25%

A, H, M, O, R: 1-Itemset
OR, MR, MO, HO, HM, AM: 2-Itemsets
HMO, MOR: 3-Itemsets

A		AH, AM, AO, AR
H		HM, HO, HR
M	⟶	MO, MR
O		OR
R		OR
MR		
MO	⟶	HMO, MOR
HO		
HM		
AM		

100%Confidence:
A->M H->A HO->M

Figure 1: Cluster 1 with 25% Support

CLUSTER 2:

H, M: 1-Itemset
HM: 2-Itemsets

H		
M	⟶	HM

90 % CONFIDENCE:
H -> M

Figure 2: Cluster 2 with 25% Support

2.4 Pattern discovery

(a) FP Growth: The proposed system is frequent pattern tree structure, which is used to develop an efficient FP Tree based mining method. FP Growth algorithm generates frequent item sets from FP-Tree by traversing in bottom up fashion. [4]

The major steps of FP-Growth algorithm:

Step 1: Construct the conditional pattern base for each item set in the header table. Then start at the

bottom of frequent item header table in the FP-Tree.

Step 2: For each pattern base.

- Accumulate the count for each item in the base
- Construct the conditional FP-Tree for the frequent item sets of the pattern base

Step 3: Recursively mine the conditional FP-Tree.

Algorithm 2 (FP-growth: Mining frequent patterns with FP-tree by pattern fragment growth).

Input: A database DB, represented by FP-tree constructed according to Apriori, and
A minimum support threshold ξ.
Output: The complete set of frequent patterns.
Method: call FP-growth (FP-tree, null).
Procedure FP-growth (Tree, α)
{
If Tree contains a single prefix path
Then
{
Let P be the single prefix-path part of Tree;
Let Q be the multipath part with the top branching node replaced by a null root;
For each combination (denoted as β) of the nodes in the path P do
Generate pattern β ∪ α with support = minimum support of nodes in β;
Let freq pattern set (P) be the set of patterns so generated;
}
Else let Q is Tree;
For each item ai in Q does
{
Construct ß's conditional pattern-base and then ß's conditional FP-tree Tree β;
If Treeβ ≠ Ø
Then call FP-growth (Treeβ, β);
Let freq pattern set (Q) be the set of patterns so generated;
}

}

FP Tree:

TID	ITEMSET
T100	K O N M E Y
T200	K O N D E Y
T300	M A K E
T400	M I C K Y
T500	C O O K I E

STEP 1: Find the support count of each item

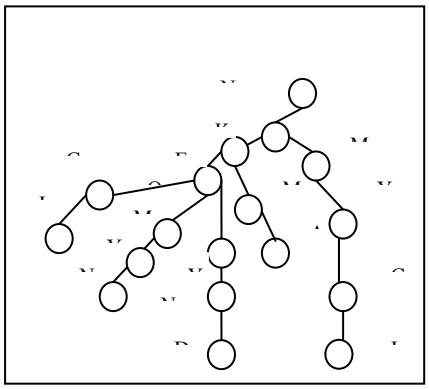
ITEMSET	SUPPORT
A	1
C	2
D	1
E	4
K	5
I	2
M	3
N	2
O	4
Y	3

STEP 2: items in descending order

ITEMSET	SUPPORT
K	5
E	4
O	4
M	3
Y	3
C	2
N	2
I	2
A	1
D	1

(FP tree is generated from step 3 to step 7)

STEP 7:
T500 : C O O K I E
Lorder: K E O O C I



2.5 Pattern Analysis

The pattern analysis is used to find frequent pattern in web usage mining, in this phase the most frequent pattern are mined.

3 RESULTS

In this paper the performance of Apriori and FP Growth Algorithm are compared. The Graph represents the performance analysis of Apriori and FP Growth algorithm. The result of this graph is that FP Growth Algorithm is more efficient because the most frequent item sets are mined compared toApriori Algorithm.

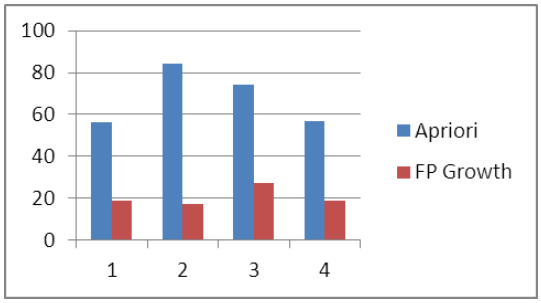


Figure 3: Performance Analysis of Apriori and FP Growth

4 CONCLUSIONS

The data from the web log files are preprocessed and the preprocessed data are stored in the database. The frequent pattern mining algorithm is applied to that data to get a most frequent pattern from the web log files.

5.REFERENCES

[1] Ashok Kumar D.Loraine Charlet Annie M.C., "web log mining using K-Apriori Algorithm", volume 41, March -2012,

[2] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, Hua Zhu, "Mining Access Pattern Efficient from Web Logs"

[3] B.Santhosh Kumar, K.V. Rukmani," Implementation of Web Usage Mining Using Apriori and FP-Growth Algorithms", volume: 01, Issue: 06, Pages: 400-404(2010)

[4] J.Han and Kamber,"Data Mining: Concepts and Techniques", Morgan Kaufman Publishers, 2000

[5] Jiawei Han, Ian Pei, Yiwen Tin, Runying Mao, "Mining Frequent Pattern without Candidate Generation: A Frequent Pattern Tree Approach", Volume-8

[6] Harish Kumar and Anil Kumar," Clustering Algorithm Employ in Web Usage Mining: An Overview", INDIA Com publication, Edition 2011.

[7] Rajan Chattamvelli, "Data Mining Methods",

[8] Narosa publications, Edition 2009.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Biometric Finger print reader using Edge Detection

Sumaya Nahar E.,
Department of Computer Science
Providence College for Women, Coonoor.

ABSTRACT

Edge Detection is one of the fundamental concepts of image processing mainly in the areas of feature detection and feature extraction. An image is an artifact that depicts or records the visual perception. In order to study the image and its characteristics, it is necessary that the edges of the objects in the image or the edges of the entire image be detected. The existing Canny Edge Detection is a traditional method with only detection of edges. Noises present in an image are also considered to be edges. In the proposed method noise is reduced and only the edges are detected. Classification of the noise based on their frequency helps in the assistance of noise reduction. One of the applications of Edge Detection, the Biometric Fingerprint Reader is studied and implemented for better understanding of Edge Detection in real time. Implementing this method in the Biometric Fingerprint Reader enhances its performance and the storage capacity is intensified.

Keywords: methods, edge, fingerprint, image, pattern

1) Introduction

An image is a two dimensional function f(x,y) with spatial coordinates x and y. The amplitude of f at any point is the intensity of the image at that point. When the values are all finite, discrete quantities, the image is classified as a digital image. A digital image is composed of a number of elements with its own location and value, introducing them to be pixels.

Digital Image Processing is a concept that refers to the processing of the Digital Images using a Digital Computer. Digital Image Processing has a number of advantages over analog image processing. There is a wide range of algorithms applied in Digital Image Processing than in analog processing. Since it provides higher performance at a simple task, it is considered more reliable than the analog methods. Digital Image Processing, a subcategory or field of digital signal processing is a process that uses computer algorithms to perform image processing on digital images. Digital image processing has many advantages over analog image processing in its efficiency and intensity.

Edge Detection is a process that involves the identification of the sharp changes in the color or orientation of an image. Various methods and algorithms used for the purpose of Edge Detection including the Classical, Laplacian of Gaussian, Machine Learning, multi resolution, Non-Linear, Statistical, Wavelet, Contextual, line and color based Detection operators are studied in the project.

2) Edge Detection Techniques

2.1) Gradient and Magnitude based Edge Detection

Edge Detection based on computing the gradient for the pixels and detecting the local maxima to localize the step edges. It is a basic method for the detection of the edges and their orientation. The technique being simple is based radically on the discrete differential operators.

The Sobel edge detector is a simple and an effective operator but is disturbed by noise easily and is not useful in detecting the outermost edges clearly. The Prewitt operator does not provide the smoothing provided by the Sobel operator. The major drawback seen in these methods is that it is sensitive to noise and is not accurate [3][4].

The Robert Cross Edge Detector is based on the 2D spatial gradient measurement of the image. The Edge Detection is performed by the high spatial frequencies. The magnitude of the spatial gradient of the input image for each different pixel is provided as the output for each pixel, a gray scale image. The convolution kernel used is as shown in the Figure1. Robert proposed the equation

1	2	1
0	0	0
-1	-2	-1

-1	0	1
-2	0	2
-1	0	1

Figure1.*Convolution kernels for Robert Edge detection*

The Sobel method is similar to the Roberts operator. It finds the approximate absolute gradient magnitude at each point. Here the operator consists of 3x3 convolution kernels. These kernels are designed to respond maximally to edges running vertically and horizontally relative to the pixel grid, one kernel for each of the two perpendicular orientations [5].

The kernels can be applied separately to the input image, to produce separate measurements of the gradient component in each orientation. These can then be combined together to find the absolute magnitude of the gradient at each point and the orientation of that gradient. One kernel is the other rotated by 90 degree. Finally, the gradient magnitude is thresholded.

The Prewitt edge detection technique was given to overcome the problem faced in Sobel Edge Detection. The operator adds a vector value in order to provide smoothing. This operator or algorithm provides a way to estimate both the magnitude and the orientation of an edge. It involves both the horizontal and the vertical edge positions. It is limited to 8 possibilities of orientation [5]. The convolution mask with the largest module is selected.

2.2) The optimal edge Detection techniques (Gaussian Method)

The method by Marr- Hilderth experimented the concept, that an edge corresponds to a change in the image orientation, thus informing that it is much intense to find the zero crossing position than the maxima. Due to the use of the Laplacian filter, the corners and curves are not detected accurately. To overcome these disadvantages the improved methods, the Canny method of the optimal Edge Detection technique is used. The traditional Canny operator analysed the product with the Gaussian first order derivative to identify the boundary.

The major improvement is seen in the noise reduction. A 2D Gaussian function is used to convolve the image. The Laplacian is computed and the pixels with zero crossing are considered as edges. Image pyramid is a method of applying repeated smoothing. The pyramid is based on two filters the low pass filter and the high pass filter [5].

The **Canny edge detector** is an edge detection operator that uses a multi-stage algorithm to detect a wide range of edges in images. Canny's aim was to discover the optimal edge detection algorithm. An "optimal" edge detector means to obtain *good detection* – the algorithm should mark as many real edges in the image as possible, *good localization* – edges marked should be as close as possible to the edge in the real image and *minimal response* – a given edge in the image should only be marked once, and where possible, image noise should not create false edges. To satisfy these requirements Canny used the calculus of variations – a technique which finds the function which optimizes a given [3][4].

The optimal function in Canny's detector is described by the sum of four exponential terms, but it can be approximated by the first derivative of a Gaussian. Because the Canny edge detector is susceptible to noise present in raw unprocessed image data, it uses a filter based on a Gaussian, where the raw image is convolved with a Gaussian filter. The result is a slightly blurred version of the original which is not affected by a single noisy pixel to any significant degree. An edge in an image may point in a variety of directions, so the Canny algorithm uses four filters to detect horizontal, vertical and diagonal edges in the blurred image. The edge detection operator (Roberts, Prewitt, Sobel for example) returns a value for the first derivative in the horizontal direction (G_x) and the vertical direction (G_y). Non-maximum suppression is an edge thinning technique. Given estimates of the image gradients, a search is carried out to determine if the gradient magnitude assumes a local maximum in the gradient direction. In many implementations, the algorithm categorizes the continuous gradient directions into a small set of discrete directions, and then moves a 3x3 filter over the output of the previous step (that is, the edge strength and gradient directions). At every pixel, it suppresses the edge strength of the center pixel (by setting its value to 0) if its magnitude is not greater than the magnitude of the two neighbors in the gradient direction.

Large intensity gradients are more likely to correspond to edges than small intensity gradients. It is in most cases impossible to specify a threshold at which a given intensity gradient

switches from corresponding to an edge into not doing so. Therefore Canny uses thresholding with hysteresis.

Thresholding with hysteresis requires two thresholds – high and low. Making the assumption that important edges should be along continuous curves in the image allows to follow a faint section of a given line and to discard a few noisy pixels that do not constitute a line but have produced large gradients. Therefore beginning by, applying a high threshold. This marks out the edges that are fairly sure to be genuine. Starting from these, using the directional information derived earlier, edges can be traced through the image. While tracing an edge, the lower threshold is applied, allowing tracing faint sections of edges as long as the starting point is found. Once this process is complete a binary image where each pixel is marked as either an edge pixel or a non-edge pixel is obtained.

A more refined approach to obtain edges with sub-pixel accuracy is by using the approach of differential edge detection, where the requirement of non-maximum suppression is formulated in terms of second- and third-order derivatives computed from a scale space representation. The proposed method has increased the signal to noise ratio thus increase the Edges being detected. This technique of Edge Detection is used in the proposed method implementation in the Biometric Finger Print reader.

3) A simple application of Edge Detection

Biometrics is an automated method of recognizing a person based on a physiological or behavioral characteristic. Among the features measured are face, fingerprints, hand geometry, handwriting, iris, retinal, vein, and voice. Biometric data are separate and distinct from personal information. Biometric templates cannot be reverse-engineered to recreate personal information and they cannot be stolen and used to access personal information.

Using a unique, physical attribute of an individual's body, such as fingerprint or iris, to effortlessly identify and verify that the person is the person who was intended. It is the best and easiest solution in the market today. That is the simple truth and power of Biometrics Technology. Although biometric technology has been around for many years, modern advances in this emerging technology, coupled with big reductions in cost, make biometrics readily available and affordable to consumers, small business owner, larger corporations and public sector agencies.

Fingerprint readers, or scanners, are the most popular and most common form of biometric security devices used. Long gone are the huge bulky fingerprint scanners; now a fingerprint scanning device can be small enough to be incorporated into a laptop for security.

A Biometric Fingerprint Scanner system has two basic jobs -- it needs to get an image of the finger, and it needs to determine whether the pattern of ridges and valleys in this image matches the pattern of ridges and valleys in pre-scanned images. Fingerprint scanning essentially provides an identification of a person based on the acquisition and recognition of those unique patterns and ridges in a fingerprint.

Only specific characteristics, which are unique to every fingerprint, are filtered and saved as an encrypted biometric key or mathematical representation. No image of a fingerprint is saved, only a series of numbers (a binary code), which is used for verification. The algorithm cannot be reconverted to an image, so no one can duplicate the fingerprints.

The actual fingerprint identification process will change slightly between products and systems. The basis of identification, however, is nearly the same. Standard systems are comprised of a sensor for scanning a fingerprint and a processor which stores the fingerprint database and software which compares and matches the fingerprint to the predefined database.

4) Study of finger prints

The analysis of fingerprints for matching purposes generally requires the comparison of several features of the print pattern. These include patterns, which are aggregate characteristics of ridges, and minutia points, which are unique features found within the patterns. It is also necessary to know the structure and properties of human skin in order to successfully employ some of the imaging technologies [6].

The three basic patterns of fingerprint ridges are the arch, loop, and whorl: (The patterns are shown in *Figure1)*

- *arch:* The ridges enter from one side of the finger, rise in the center forming an arc, and then exit the other side of the finger.
- *loop:* The ridges enter from one side of a finger, form a curve, and then exit on that same side.

- *whorl:* Ridges form circularly around a central point on the finger.



Figure 1

5) Existing - Pattern Matching in Finger prints

Pattern based algorithms compare the basic fingerprint patterns (arch, whorl, and loop) between a previously stored template and a candidate fingerprint [1]. This requires that the images be aligned in the same orientation. To do this, the algorithm finds a central point in the fingerprint image and centers on that. In a pattern-based algorithm, the template contains the type, size, and orientation of patterns within the aligned fingerprint image. The minutia features are extracted from the image and stored. The candidate fingerprint image is graphically compared with the template to determine the degree to which they match [2]. The working of the pattern matching based finger print reader is depicted in *Figure2*.

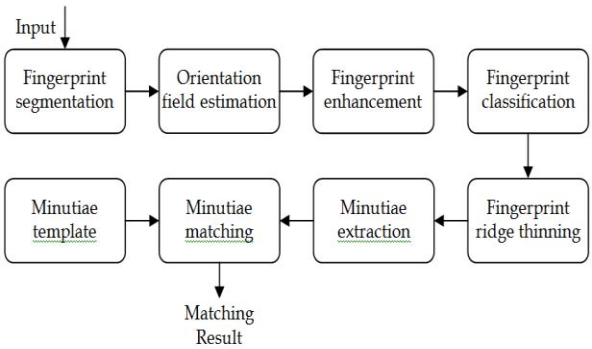


Figure2 Pattern matching in Finger print reader

6) Proposed- Edge detection in Finger prints

When considering the Biometric fingerprint reader it has made its appearance long back in the scenario. But the concept used in almost all the fingerprint reader is "Pattern matching" where the patterns of the existing image are assigned numbers and its match is found. Here both the image and the number are to be stored. The concept of Edge Detection is used in the proposed technique to make this process more efficient. Here the edges in the image are discovered and are mathematically formulated to produce a value that is stored in the database for comparison.

The Canny Edge Detection operator described in the optimal Edge Detection technique is used as bases in the proposed method. The Characteristics such as good localization, good detection and minimal response with increase in signal to noise ratio, of the operator helps detect the edges with perfection. The proposed method is implemented as shown in *Figure3*.

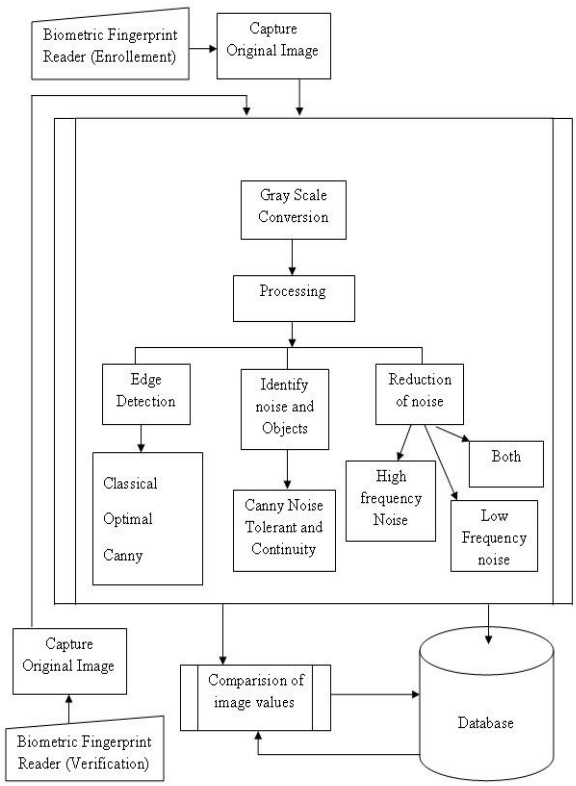


Figure3 Working of the proposed method in the Biometric Fingerprint Reader

7) Comparison of the Outputs obtained in Edge Detection



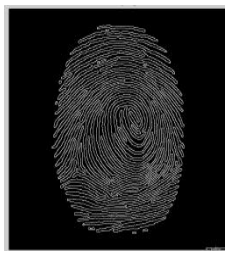
Original Image



Noisy Image



Traditional Canny



Noise Tolerant Canny



Proposed

8) Conclusion

The concept of Edge Detection in Digital Image Processing has enhanced and spread the applications of Digital Image Processing. Edge Detection is almost used in a number of application including satellite imagery, vehicle recognition, etc. and along with other techniques like wavelets, neural networks, resolution, anisotropic analysis, multi-scale analysis and others it is implemented in different fields including medical and industrial fields. Implementing the proposed Edge Detection technique in the Biometric Fingerprint reader provides accuracy and the amount of information to be stored and retrieved from the database becomes very less, thus, improving its performance.

References

1) Ravi. J, K. B. Raja, and Venugopal. K.R "Fingerprint Recognition Using Minutia Score Matching" International Journal of Engineering Science and Technology Vol.1 (2), 2009, 35-42, ISSN: 0975-5462.
2) Anil K. Jain, Jianjiang Feng, and Karthick Nandakumar, "Fingerprint Matching" IEEE Computer Society 0018-9162.
3) Raman Maini and Dr. Himanshu Aggarwal, "Study and Comparison of Various Image Edge Detection Techniques" International Journal of Image Processing (IJIP), Volume (3) : Issue (1)
4) G.T. Shrivakshan and Dr.C. Chandrasekar "A Comparison of various Edge Detection Techniques used in Image Processing" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 1, September 2012. ISSN (Online): 1694-0814.
5) M. Kalpana, G. Kishorebabu, and K.Sujat, "Extraction of Edge Detection Using Digital Image Processing Techniques" International Journal Of Computational Engineering Research (ijceronline.com) Vol. 2 Issue.5 Issn 2250-3005September| 2012 .
6) Sharath Pankanti, Salil Prabhakar, and Anil K. Jain, "On the Individuality of Fingerprints" IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, NO. 8, August 2002

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A Survey on Dimensionality Reduction Techniques for Document Clustering

[1] C.Rathika [2] K.Shankri

[1 & 2] Assistant Professor,
Department of Computer Science, Sri Ramakrishna College for Women,
Coimbatore.

Abstract

Document clustering has been used for better document retrieval, document browsing, and text mining in a digital library. So, it is a major issue to retrieve relevant documents from the larger database, to improve the retrieval performance either via cluster-driven dimensionality reduction, term-weighting or query expansion. Document Clustering is the method of Segmenting a scrupulous Collection of text into sub groups including content based similar ones. Now a day's all paper documents in Electronic form because of quick access and smaller storage. The challenging problems of document clustering are big volume, High dimensionality and Complex Semantics. This paper mainly focus on High dimensional Reduction techniques and discuss the key methods used for text clustering and their relative advantages and also a detailed survey of text clustering with High Dimensional Reduction and the Key challenges of the clustering problem.

Index Terms - Clustering, Document Clustering, High Dimensional reduction

I. INTRODUCTION

The term Knowledge Discovery from data (KDD) refers to the automated process of knowledge discovery from databases. The process of KDD is comprised of many steps namely data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and Knowledge representation. Data mining is a step in the whole development of knowledge discovery which can be explained as a process of extracting or mining knowledge from large amounts of data [1]. Data mining is a form of knowledge discovery essential for solving problems in a specific domain. Data mining can also be explained as the non trivial process that automatically collects the useful hidden information from the data and is taken on as forms of rule, concept, pattern and so on [2]. The knowledge extracted from data mining, allows the user to find interesting patterns and regularities deeply buried in the data to help in the process of decision making. The data mining tasks can be broadly classified in two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. According to different goals, the mining task can be mainly divided into four types: class/concept description, association analysis, classification or prediction and clustering analysis [3].

II. CLUSTERING

Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering. A general definition of clustering states a number of objects or individuals, each of which is described by a set of numerical measures, devise a classification scheme for grouping the objects into a number of classes such that objects within classes are similar in some respect and unlike those from other classes. The number of classes and the characteristics of each class are to be determined. Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters[1].

III. TYPES OF CLUSTERING METHODS

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired. In general,

clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap. These methods are sometimes divided into *partitioning* methods, in which the classes are mutually exclusive, and the less common *clumping* method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar, however the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be divided into *agglomerative* or *divisive methods*. Some of the important Data Clustering Methods are described below.

A. Partitioning Methods

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

Single Pass: A very simple partition method, the single pass method creates a partitioned dataset as follows:

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient.
3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

As its name implies, this method requires only one pass through the dataset; the time requirements are typically of order *O(NlogN)* for order *O(logN)* clusters. This makes it a very efficient clustering method for a serial processor. A disadvantage is that the resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run

B. Hierarchical Agglomerative methods

Hierarchical method proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The clustering methods differ in the rule by which it is decided which two small clusters are merged or which large cluster is split. The end result of the algorithm is a tree of clusters called a dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level a clustering of the data items into disjoint groups is obtained.

The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The name *single link* thus refers to the joining of pairs of clusters by the single shortest link between them[9].

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters[3].

The group average method relies on the average value of the pair wise within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods. Since all objects in a cluster contribute to the inter –cluster similarity, each object is , on average more like every other member of its own cluster then the objects in any other cluster[4].

In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. Now when a query comes regarding a typical word then instead of checking the entire database, only that cluster is scanned which has that word in the list of its key words and the result is given. The order of the documents received in the result is dependent on the number of times that key word appears in the document[9].

IV. DOCUMENT CLUSTERING

The goal of a document clustering method is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an

appropriate distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of document clustering. The huge selection of documents makes it almost impossible to create a general algorithm which can work best in case of all kinds of datasets. Document clustering is a technique for identifying clusters or groups of documents which share some common features or have overlapping content. These groupings of documents can be useful in document retrieval. Document clustering has been used in experimental Information Retrieval (IR) systems for decades. As stated by [4], [5].

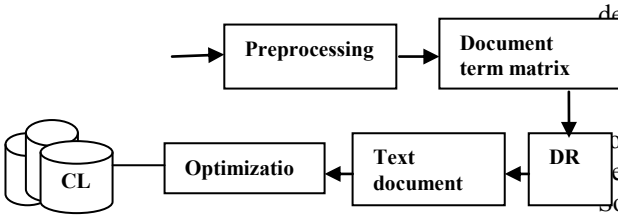


Figure 1:Flow diagram for Representing basics steps in Clustering

Data clustering algorithms can be broadly classified into following categories: *Partitioning methods, Hierarchical methods, Density-based methods, Grid-based methods, Model-based methods, Frequent pattern-based clustering, Constraint-based clustering.* A major Characteristics of document clustering algorithms is the high dimensionality of the feature space, which imposes a big challenge to the performance of clustering algorithms. Next challenge is the features may be redundant or irrelevant and some may even misguide the clustering result [6]. Some of the important Document Clustering techniques are described below:

Partitional clustering the algorithm creates a set of data non-overlapping subsets (clusters) such that each data object is in exactly one subset. These approaches require selecting a value for the desired number of clusters to be generated. A few popular heuristic clustering methods are k- means and a variant of k-means-bisecting k-means, k-medoids, PAM (Kaufman and Rousseeuw, 1987), CLARA (Kaufmann and Rousseeuw, 1990), CLARANS (Ng and Han, 1994) etc.

Hierarchical clustering the algorithm creates a nested set of clusters that are organized as a tree. Such hierarchical algorithms can be agglomerative or divisive. Agglomerative algorithms, also called the bottom-up algorithms, initially treat each object as a separate cluster and successively merge the couple of clusters that are close to one another to create new clusters until all of the clusters are merged into one. Divisive algorithms, also called the top-down algorithms, proceed with all of the objects in the same cluster and in each successive iteration a cluster is split up using a flat clustering algorithm recursively until each object is in its own singleton cluster. The popular hierarchical methods are BIRCH [7], ROCK [8], Chamelon [9] and UPGMA. An experimental study of hierarchical and partitional clustering algorithms was done by [9] and proved that bisecting kmeans technique works better than the standard kmeans approach and the hierarchical approaches.

Density-based clustering methods group the data objects with arbitrary shapes. Clustering is done according to a density (number of objects), (i.e.) density-based connectivity. The popular density-based methods are DBSCAN and its extension, OPTICS and DENCLUE [10].

Grid-based clustering methods use multitier resolution grid structure to cluster the data objects. The benefit of this method is its speed in processing time. Some examples include STING, Wave Cluster.

Model-based methods use a model for each cluster and determine the fit of the data to the given model. It is also used to automatically determine the number of clusters. Expectation-Maximization, COBWEB and SOM (Self-Organizing Map) are typical examples of model-based methods. Frequent pattern-based clustering uses patterns which are extracted from subsets of dimensions, to group the data objects. An example of this method is pCluster.

Constraint-based clustering methods perform clustering based on the user-specified or application-specific constraints. It imposes user's constraints on clustering such as user's requirement or explains properties of the required clustering results.

V. CHALLENGES IN DOCUMENT CLUSTERING

- High Dimensionality: High dimensionality greatly affects the scalability and efficiency of many existing clustering algorithms.
- High Volume of Data: In text mining processing of data about ten thousands to hundred thousands documents are involved .
- Consistently High accuracy: A few existing algorithms only work fine for certain type of document sets ,but may not perform well in some others.
- Consequential Cluster Description: this is important for the end user. The resulting hierarchy should facilitate browsing.
- Select proper features of the document that should be used for clustering.
- Select a suitable similarity measures between documents.

- Selecting a correct clustering method utilizing the above similarity measures.
- Implementing the clustering algorithm in a efficient way that makes it feasible in terms of required memory and CPU resources.
- Finding ways of assessing the quality of the performed clustering.

VI. DIMENSIONALITY REDUCTION TECHNIQUES

In machine learning and statistics **dimensionality reduction** or **dimension reduction** is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction. Both are the major techniques of dimensionality reduction. Feature Selection is a process that chooses an optimal subset of features according to a objective function. The main objective is to reduce dimensionality and remove noise and improve mining performance like Speed of learning, Predictive accuracy and Simplicity and comprehensibility of mined results. Feature reduction refers to the mapping of the original high-dimensional data onto a lower dimensional space. Dimension reduction is commonly defined as the process of mapping high-dimensional data to a lower-dimensional embedding. Applications of dimension reduction include, but are not limited to, filtering, compression, regression, classification, feature analysis, and visualization. The aim is to provide an overview of basic approaches, as well as to review select state-of-the-art methods.

As one of the first dimension reduction techniques discussed in the literature, Principal Components Analysis (PCA) conveys distance relationships of the data by orthogonally projecting it on a linear subspace of target dimensionality. In this specific subspace, the orthogonally projected data has maximal variance. Thereby, PCA defines a "faithful" approximation as one that captures the data's variance in an optimal way. It has been shown [11] Y. Koren and L. Carmel that by the maximization of variance, PCA also minimizes the least squares error for Euclidean distances in data and target space. Remarkably, PCA achieves this through a computationally efficient linear transformation. The resulting projection is a genuine view that does not distort the data. The only major drawback of PCA is that, due to its linear nature, it does not capture non-linear data well. Besides its broad applicability to visualization, PCA may be used for many more tasks.For example, a prominent gap in the eigen value spectra gives an upper bound for the intrinsic dimensionality of the data. Therefore, it is often used for filtering Gaussian noise or for reducing data size and computation time. PCA is a well-established technique with an extensive history. As such, many variants exist and more information can be found, for example, in [12. T. Jolliffe] or [13. B. F. Manly].

Metric Multidimensional Scaling (MDS)[14. W.S. Torgerson], also known as classical MDS, is a well-established approach that uses projection to map high-dimensional points to a linear subspace of lower dimensionality. The technique is often motivated by its goal to preserve pair wise distances in this mapping. As such, metric MDS defines a faithful approximation as one that captures pair wise distance relationships in an optimal way; more precisely, inner product relations. Metric MDS finds an optimal (least squares) linear fit to the given pair wise distances, assuming the distance used is metric.

Kernel PCA[15]. Bernhard Schölkopf, Alex J. Smola, and Klaus-Robert Müller. is considered a variant of PCA and metric MDS (due to their duality) that is capable of depicting non-linear data. Although distance relationships along a non-linear pattern are unknown, Kernel PCA is based on two assumptions that makes the application of (linear) PCA to non-linear data possible. The first assumption is that in the space of the data's underlying features, the data is linear. The second assumption is that there is a function that approximates the inner product of data points in this feature space. This function is called a kernel and the utilization of a non-linear kernel in a linear setting to capture non-linear data structure is known as the "kernel trick". It is not surprising that the bottleneck of Kernel PCA is finding the "right" kernel. Since distance relationships along the possibly non-linear sub-structures of the data are, in general, a-priori unknown, the definition of a suitable kernel requires explicit knowledge about the data. If this knowledge is not given, methods are better suited that determine distance relationships along non-linear data structures in an unsupervised data-driven manner. This is the concept of manifold learning.

The ability of metric MDS to map data relationships from a dissimilarity matrix is based on the key assumption that dissimilarities are approximate squared metric distances. As for all spectral methods, this allows for the computation of a global optimal projection. However, this also limits its application and prohibits non-metric scenarios, for example, stemming from psychometric research where metric postulates do not hold. Instead of this eigen decomposition approach, the idea of non-metric Multidimensional Scaling is to directly minimize the mapping error with respect to a given non-metric dissimilarity matrix and possibly some weighting

thereof. Unfortunately, due to the non-metric nature, the resulting stress function is non-convex and optimization thereof is prone to local minima. The exact embedding of non-metric dissimilarities in a metric target space is impossible. A well-known approach to non-metric MDS is the Shepard-Kruskal algorithm[16.] J. Kruskal. At its core is a twofold optimization process that optimizes the goodness of fit with regard to the non-metric input. First, an optimal monotonic transformation of the non-metric dissimilarities to metric distances is found that preserves the rank-order of nonmetric inputs. After the optimization of the rank-order distances, the output configuration is further improved iteratively, balancing both stress and monotonicity.MDS is in all respects a hard non-convex optimization problem. Using a good initialization is therefore important. Numerous variants of MDS exist and many other methods are closely related, like Sammon's mapping[17]. Sam T. Roweis and Lawrence K. Saul. Especially multi-level approaches have substantially increased performance[18]. Stephen Ingram, Tamara Munzner, and Marc Olano. Glimmer. For an overview, reference [19]. Borg and P.J.F. Groenen. is helpful.

Instead of learning the embedding directly in target space, Isomap[20]. J.B. Tenenbaum, V. Silva, and J.C. Langford. attempts to explicitly model non-linear proximity relationships in terms of geodesic distances. As such, it can be viewed as a variant of metric MDS to model non-linear data using its (metric) geodesic distances. In order to retrieve these distances, a global graph-based optimization approach is utilized. Geodesic distances are learned by linearly approximating the non-linear manifold. There by, a network of undirected neighborhood graphs is constructed in which each data point is a node and has edges to its neighbors that are weighted by the points' dissimilarity. The weights represent the local approximation of geodesic distances on the manifold. From these graphs, a square geodesic distance matrix is computed which is used for the metric MDS projection. One problem of Isomap is that after double-centering of the geodesic distances, the Gram matrix of inner products is not guaranteed to be positive semidefinite. One variant that solves this issue is Maximum Variance Unfolding (MVU)[21]. Kilian Q. Weinberger and Lawrence K. Saul.. The underlying idea behind MVU is to unfold the manifold under the constraint that local distances between neighboring points are preserved. This is optimized with respect to maximum variance.

In contrast to modeling a manifold by global geodesic distance relationships, LLE[22]. Sam T. Roweis and Lawrence K. Saul. models the manifold by extracting its local intrinsic geometry. Thereby, LLE follows a local graph based approach. The basic idea of LLE is based on the linear approximation of all data points (in complex non-linear structures) by a convex linear combination of its neighborhood. In contrast to Isomap, LLE models nearest neighbors by directed graphs which leads to a more suitable approximation. With these local relationships, LLE constructs a set of global equations for the projection to target space.

SDE can be seen as a variation on kernel PCA, in which the kernel matrix is also learned from the data. This is in contrast with classical kernel PCA which chooses a kernel function a priori. To derive SDE, Weinberger and Saul formulated the problem of learning the kernel matrix as an instance of semidefinite programming.

VII. CONCLUSION

Research on dimension reduction continues at a rapid pace. This survey provides an introduction to the main concepts of dimension reduction and reduce the complexity of high dimensional datasets. The paper describes a survey of dimensionality reduction techniques for document clustering h objective functions that can be optimized well. Different sampling techniques should be able to provide even more fruitful success in reducing the computational cost without sacrificing its performance.

REFERENCES

1. J. Han and M. Kamber, *Data mining concepts and techniques*, Morgan Kaufmann, San Francisco 2006
2. T.J. Shan, H. Wei and Q. Yan, "Application of genetic algorithm in data mining", *1st Int Work EducTechnol Comput Sci, IEEE 2*, 2009, pp. 353- 356
3. Z.Z. Shi, *Knowledge discovery*, Tsinghua University Press, Beijing, 2001.
4. Jiawei han & Michelin Kamber ,(2010),‖Data mining concepts and techniques‖,Elsevier.
5. Tan,Steinbach & kumar, (2009), ―Introduction to data mining‖, Pearson education .
6. Fabrizio Sebastiani "Machine Learning in Automated Text Categorization" ACM Computing Surveys, Vol. 34, No. 1, March 2002
7. Zhang.T., Raghu Ramakrishnan & Livny.M.,(1996), ―Birch: An Efficient Data Clustering Method for very Large Databases‖, In Proceedings of the ACM SIGMOD

international Conference on Management of Data, pp. 103-114.

8. S. Guha, R. Rastogi, and K. Shim,(1999), ―ROCK: A robust clustering algorithm for categorical attributes‖, International Conference on Data Engineering (ICDE―99), pp. 512-521.
9. Karypis.G, Eui-Hong Han & Kumar.V., (1999), ―Chameleon: A Hierarchical Clustering Algorithm using Dynamic Modelling‖, IEEE Computer, Vol. 32, No.8, pp. 68-75.
10. Jiawei han & Michelin Kamber ,(2010),‖Data mining concepts and techniques‖,Elsevier.
11. Y. Koren and L. Carmel. Robust linear dimensionality reduction. *Visualization and ComputerGraphics, IEEE Transactions on*, 10(4):459 –470, jul. 2004
12. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, 2002.
13. B. F. Manly. *Multivariate statistical methods: a primer*. Chapman & Hall, Ltd., London,UK, UK, 1986.
14. W.S. Torgerson. *Theory and methods of scaling*. Wiley, 1958.
15. Bernhard Schölkopf, Alex J. Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.
16. J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1964.
17. Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.
18. Stephen Ingram, Tamara Munzner, and Marc Olano. Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics*, 15:249–261, 2009.
19. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*.Springer, 2005
20. J.B. Tenenbaum, V. Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
21. Kilian Q. Weinberger and Lawrence K. Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI*. AAAI Press, 2006.
22. Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

MOBILE AD HOC NETWORKING USING WIRELESS MOBILE

[1]T.Sangeetha [2]P.Ramesh

[1]M.Phil Research scholar, Department Of Computer Science, Kongu Arts And Science College

[2]Assistant Professor And Head Of The Department, Department Of Computer Science, Kongu Arts And Science College, Erode.

ABSTRACT

Mobile ad hoc networks (MANETs) represent complex distributed systems that comprise wireless mobile nodes that can freely and dynamically self-organize into arbitrary and temporary, ‘‘ad-hoc’’ network topologies, allowing people and devices to seamlessly internetwork in areas with no pre-existing communication infrastructure, e.g., disaster recovery environments. Traditionally, tactical networks have been the only communication networking application that followed the adhoc paradigm. Recently it is used to a traveller with portable computers can surf the internet from airports, railway stations, Starbucks and other public locations; tourists can use Global Positioning System terminals installed inside rental cars to locate driving maps and tourist attractions, researchers can exchange files and the introduction of new technologies such as the Bluetooth, IEEE 802.11 and Hyperlink are helping enable eventual commercial MANET deployments outside the military domain. These recent evolutions have been generating a renewed and growing interest in the research and development of MANET. This paper attempts to provide a comprehensive overview of this dynamic field. It first explains the important role that mobile ad hoc networks play in the evolution of future wireless technologies. Then, it reviews the latest research activities in these areas, including summary of MANET_s characteristics, capabilities, applications, and design constraints. The paper concludes by presenting a set of challenges and problems requiring further research in the future.

Keywords: MAC; Routing; Energy saving; Security; Performance evaluation

1. Introduction

The mobile computing and communication devices (e.g., cell phones, laptops, handheld digital devices, personal digital assistants, or wearable computers) is driving a revolutionary change in our information society. We are moving from the Personal Computer age (i.e., alone computing device per person) to the Ubiquitous Computing age in which a user utilizes, at the same time, several electronic platforms through which he can access all the required information whenever and wherever needed . The nature of ubiquitous devices makes wireless network easiest solution for their interconnection and, consequence, the wireless arena has been experiencing exponential growth in the past decade. Mobile users can use their cellular phone to checked-mail, browse internet; traveler with portable computers can surf the internet from airports, railwastations, Starbucks and other public locations; tourists can use Global Positioning SystemGPS) terminals installed inside rental cars to locate driving maps and tourist attractions, researchers can exchange files and other information by connecting portable computers via wirelessLANs while attending conferences. With this trend, we can expect the total number of mobile Internet users soon to exceed that of the fixed-line Internet users. Among all the applications and services run by mobile devices, network connections and corresponding data services are without doubt the most demanded service by the mobile users. As the wireless network continues to evolve, these ad hoc capabilities are expected to become more important, the technology solutions used to support more critical and significant future research and development efforts can be expected in industry and academy, alike. This paper demonstrates the impetus behind mobile ad hoc networks, and presents a representative collection of technology solutions used at the different layers of the network, in particular presenting algorithms and protocols unique to the operation and dynamic configuration of mobile adhoc networks. Mobile ad hoc network

(MANET)literature is already too expensive to be covered and analyzed in detail in this article. Hereafter, we therefore present the main research areas in the MANET literature, and inside each, survey the main research directions and open issues. Inside the ad hoc networking field, wireless sensor networks take a special role. A sensor networks composed of a large number of small sensor nodes, which are typically densely (and randomly) deployed inside the area in which phenomenon is being monitored. Wireless ad hoc networking techniques also constitute the basis for sensor networks. However, the special constraints imposed by the unique characteristics of sensing devices, and by the application requirements, make many of the solutions designed for multichip wireless networks (generally) not suitable for sensor networks [12]. This places extensive literature dedicated to sensor networks beyond the scope of this paper; however, the interested reader can find an excellent and comprehensive coverage of sensor networks in a recent survey [12].

2. 4G and ad hoc networking

A major goal toward the 4G Wireless evolutionism the providing of pervasive computing environments that can seamlessly and ubiquitously support users in accomplishing their tasks, in accessing information or communicating with other users at anytime, anywhere, and from any device . In this environment, computers get pushed further into background; computing power and network connectivity are embedded in virtually every device to bring computation to users, no matter where they are, or under what circumstances they work. These devices personalize themselves in our presence to find the information or software we need. The new trend is to help users in the tasks of everyday life by exploiting technologies and infrastructures hidden in the environment, without requiring any major change the users behavior. This new philosophy is the basis of the Ambient Intelligence concept [1]. The objective of ambient intelligence is the integration of digital devices and networks into the everyday environment, rendering accessible, through easy and ‘‘natural’’ interactions, a multitude of services and applications. Ambient intelligence places the user at the center of the information society. This view heavily relies on4G wireless and mobile communications. 4G is all about an integrated, global network, based on an open systems approach. Integrating different types of wireless networks with wire-line backbone network seamlessly and convergence of voice, multimedia and data traffic over a single IP-based core network are the main foci of 4G. With the availability of ultra-high bandwidth of up to 100 Mbps, multimedia services can be supported efficiently; ubiquitous computing is enabled with enhanced system mobility and portability support, and location-based services are all expected. Fig. 1 illustrates the networks and components within 4Gnetwork architecture. Network Integration. 4G networks are touted as hybrid broadband networks that integrate different network topologies and platforms. In Fig. 1 the overlapping of different network boundaries represents the integration of different types of networks in 4G. There are two levels of integration. First is the integration of heterogeneous wireless networks with varying transmission characteristics such as Wireless LAN, WAN, PAN, as well as mobile ad hoc networks. At the second level we find the integration of wireless networks with the fixed network backbone infrastructure, the Internet, and PSTN. This evolution is expected to greatly simplify the network and to reduce costs for maintaining separate networks, for different traffic types.

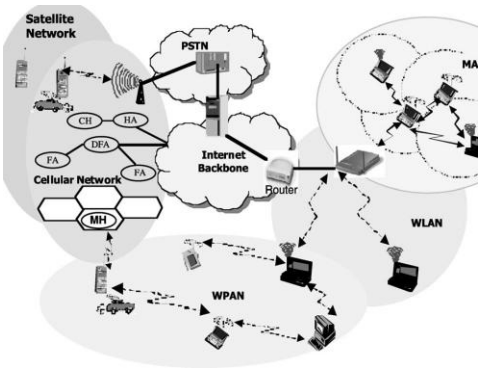


Fig. 1. 4G networks.

Ultra-High Speed and Multimedia Applications:

4G systems aim to provide ultra-high transmission speed of up to 100 Mbps, 50 times faster than those in 3G networks. This leap in provided bandwidth will enable high-bandwidth wireless services, allowing users to watch TV, listen to the music, browse Internet, access business programs, perform real-time video streaming and other multimedia-oriented applications, like E-Commerce, as if sitting in home or office.

Location Intelligence:

To support ubiquitous computing requirements, 4G terminals need to before intelligent in terms of users locations and service

needs, including recognizing and being adaptive to users changing geographical positions, as well as offering location-based services .Anytime anywhere requires intelligent use of location information, and the embedding of the information into various applications. Possible Location Based Services include finding nearest service providers, such as restaurant or cinema; searching for special offers within an areas; warning of traffic or weather situations; sending an advertisement to a specific area; searching for other users; active badge systems, etc. Outdoor, wireless applications can use GPS to obtain location information. GPS is a satellite-based system that can provide easy, accurate positioning information almost anywhere on earth. Many GPS implementations are available, including integrating GPS receiver into a mobile phone (GPS/DGPS);

3. Mobile ad hoc networks

As concluded in Section 2, ad hoc networking capabilities can become essential in delivering Overall next generation wireless network functionalities. Next, we will look at mobile hocnetworks applications from an historical perspective, and then we will focus on challenges in the MANET research activities

3.1. MANET evolution

Historically, mobile ad hoc networks have primarily been used for tactical network related applications to improve battlefield communications/survivability. The dynamic nature of military operations means that military cannot rely on access to a fixed pre-placed communication infrastructure in battlefield. In addition, by using multi-hop store-and-forward routing techniques, the radio coverage limitation is removed, which effectively enables multi-user communication within a very large geographic area. Survivable Radio Networks (SURAN) were developed by DARPA in 1983 to address main issues in PRNet, in the areas of network scalability, security, processing capability and energy management. The main objectives were to develop network algorithms to support a network that can scale to tens of thousands of nodes and withstand. Security attacks, as well as use small, low-cost, low-power radios that could support sophisticated packet radio protocols. Direct-sequence spread-spectrum, time division multiple access radio is used with data rates in the tens of kilobits per second ranges, while modified commercial Internet protocols are used for networking among nodes.

Autonomous and infrastructure-less:

MANETdoes not depend on any established infrastructure or centralized administration. Each node operates in distributed peer-to-peer mode, acts as an independent router and generates independent data. Network management has to be distributed across different nodes, which brings added difficulty in fault detection and management.

Multi-hop routing:

No default router available, every node acts as a router and forwards each other's_ packets to enable information **sharing** between mobile hosts. Dynamically changing network topologies. In mobile ad hoc networks, because nodes can move arbitrarily, the network topology, which is typically multi-hop, can change frequently and unpredictably, resulting in route changes, frequent network partitions, and possibly packet losses.

Variation in link and node capabilities:

Each node may be equipped with one or more radio interfaces that have varying transmission/receiving capabilities and operate across different frequency bands. This heterogeneity in node radio capabilities can result in possibly asymmetric links. In addition, each mobile node might have a different software/hardware configuration, resulting in variability in processing capabilities. Designing network protocols and algorithms for this heterogeneous network can be complex, requiring dynamic adaptation to the changing conditions (power and channel conditions, traffic load/distribution variations, congestion, etc.).Energy constrained operation. Because batteries carried by each mobile node have limited power supply, processing power is limited, which in turn limits services and applications that can be supported by each node. This becomes a bigger issue in mobile ad hoc networks because, as each node misacting as both an end system and a router at the same time, additional energy is required to forward packets from other nodes.

A layered approach into three main areas:

- Enabling technologies;
- Networking;
- Middleware and applications.

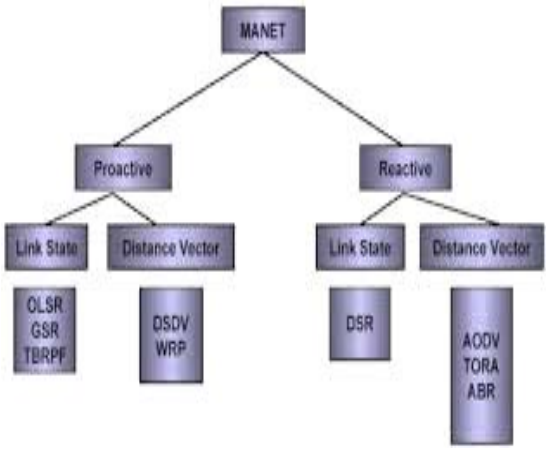


Fig.2 A simple MANET

In addition, as shown in the figure, several issues (energy management, security and cooperation, quality of service, network simulation) span all areas, and we discuss them separately. High capacity wireless technologies, etc.

3.3. Ad hoc networking research

The specific MANET issues and constraints described above pose significant challenges in ad hoc network design. A large body of research has-been accumulated to address these specific issues, and constraints.

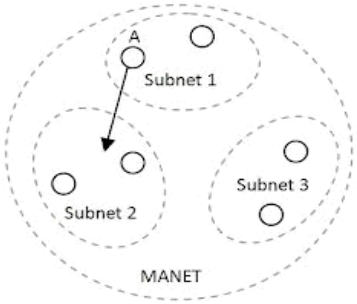


Fig. 3. A simple MANET architecture.

4. Enabling technologies

As shown in Fig. 3, we can classify ad hoc networks, depending on their coverage area, into

Several classes: Body (BAN), Personal (PAN), Local (LAN), Metropolitan (MAN) and Wide

(WAN) area networks. Wide- and Metropolitan-area ad hoc networks are mobile multi-hop wireless networks that present many challenges that are still to be solved (e.g., addressing, routing, location management, security, etc.), and their availability is not on immediate horizon. Two different approaches can be followed in the implementation of a WLAN: an infrastructure based approach or an ad hoc networking one .An infrastructure-based architecture imposes the existence of a centralized controller for each cell, often referred to as Access Point. The Access Point(AP) is normally connected to the wired network, thus providing the Internet access to mobile devices. In contrast, an ad hoc network is a peer to-peer network formed by a set of stations within the range of each other, which dynamically configure themselves to set up a temporary network. In the ad hoc configuration, no fixed controller is required, but a controller may be dynamically elected among the stations participating in the communication. The success of a network technology is connected to the development of networking products at a competitive price. A major factor in achieving this goal is the availability of appropriate networking standards.

4.1. Bluetooth

The Bluetooth technology is a de-facto standard for low-cost, short-range radio links between

Mobile PCs, mobile phones, and other portable devices . The Bluetooth Special Interest Group (SIG) releases the Bluetooth specifications. Enables wireless ad hoc communications, of voice and data between portable and/or fixed electronic devices like computers, cellulaphones, printers, and digital cameras. Due to its low-cost target, Bluetooth microchips may become embedded in virtually all consumer electronic devices in the future. As a low cost, low power solution and with industry-wide support, Bluetooth wireless technology has already started to revolutionize the personal connectivity market by providing freedom from wired connections—enabling portable links between mobile computers, mobile phones, portable handheld devices, and connectivity to the Internet. Eventually, microcellular-based Personal Area Networks will able to provide services such as real-time voice and data in a much more economical way than in existing systems.

4.2. IEEE 802.11 networks

The IEEE 802.11 standard defines two operational modes for WLANs: infrastructure-based and infrastructure-less or ad hoc. Network interface cards can be set to work in either of these modes but not in both simultaneously. Infrastructure mode resembles cellular infrastructure-based networks. It is the mode commonly used to construct the so-called Wi-Fi hotspots, i.e., to provide wireless access to the Internet. In the ad hoc mode, any station that is

within the transmission range of any other, after a synchronization

Phase ,can start communicating.

Network Integration: 4G networks are touted as hybrid broadband networks that integrate different network topologies and platforms. overlapping of different network boundaries represents the integration of different types of networks in 4G. There are two levels of integration. First is the integration of heterogeneous wireless networks with varying transmission characteristics such as Wireless LAN, WAN, PAN, as well as mobile ad hoc networks. At the second level we find the integration of wireless networks with the fixed network backbone infrastructure, the Internet, and PSTN. Much work remains to enable a seamless integration, for example that can extend IP to support mobile network devices. All IP Networks: 4G starts with the assumption that future networks will be entirely packet-switched, using protocols evolved from those in use into days Internet. An all IP-based 4G wireless network has intrinsic advantages over its predecessors.IP is compatible with, and independent of, the actual radio access technology, this means that the core4G network can be designed and evolves independently from access networks. Using IPbasedcore network also means the immediate tapping of the rich protocol suites and services already available, for example, voice and data convergence, can be supported by using readily available VoIP set of protocols such as MEGACOP,MGCP, SIP, H.323, SCTP, etc. Finally the converged all-IP wireless core networks will be packet based and support packetized voice and multimedia on top of data. This evolution is expected greatly simplify the network and to reduce costs for maintaining separate networks, for different traffic types.

No AP is required, but if one of the stations operating in the ad hoc mode has a connection also to a wired network, stations forming the ad hoc network gain wireless access to the Internet. In a MANET, the users_ mobile devices are the network, and they must cooperatively provide the functionality usually provided by the network infrastructure (e.g., routers, switches, servers). In a MANET, no infrastructure is required to enable information exchange among users_ mobile devices. We can envisage these devices as an evolution of current mobile phones, and emerging PDA_s equipped with wireless interfaces. The only external resource needed for their successful operations the bandwidth, often the (unlicensed) ISM band. Nearby terminals can communicate directly by exploiting, for example, wireless LAN Technologies. Devices that are not directly connected, communicate by forwarding their trafficvia a sequence of intermediate devices.

MANETs are gaining momentum because they help realizing network services for mobile users in areas with no pre-existing communications infrastructure, or when the use of such infrastructure requires wireless extension. Ad hoc nodes can also be connected to a fixed backbone network through a dedicated gateway device enabling IP networking services in the areas where Internet services are not available due to a lack of preinstalled infrastructure. All these advantages make ad hoc networking an attractive option in future

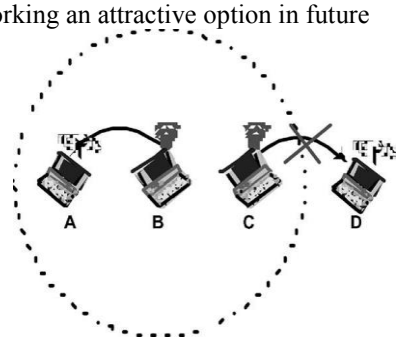


Fig. 5. Exposed-station problem.

To cope with the self-organizing, dynamic, volatile, peer-to-peer communication environment

in a MANET, most of the main functionalities of the Networking protocols (i.e., network and transport protocols in the Internet architecture) need to be re-designed. In this section we provide an outline of the main research issues in these areas, and survey the existing literature. The aim of the networking protocols is to use the one-hop transmission services provided by the enabling technologies to construct end-to-end(reliable) delivery services, from a sender to one(or more) receiver(s). To establish an end-to-end communication, the sender needs to locate the receiver inside the network. The purpose of a location service is to dynamically map the logical address of the (receiver) device to its current location in the network. Current solutions generally adopted to manage mobile terminals in infrastructure networks are generally inadequate, and new approaches have to be found. Once, a user is located, routing and forwarding algorithms must be provided to route the information through the MANET. Finally, the low reliability of communications (due to wireless communications, users_ mobility, etc.), and the possibility of network congestion require a redesign of Transport Layer mechanisms.

5.1. Location services

A Location Service answers queries about nodes_ location. In legacy mobile networks (e.g., GSM, Mobile IP), the presence of a fixed infrastructure led to the diffusion of two-tier schemes to track the position of mobile nodes. Examples are the Home Location Register/Visitor Location Register approach used in GSM networks, and the Home Agent/Foreign Agent approach for Mobile IP networks. Efficient implementations of these approaches use centralize servers. In a mobile ad hoc network, these solutions are not useful, and new approaches have to be found for mobility management [198].A simple solution to node location is based on Of course, flooding does not scale, and hence this approach is only suitable for limited size networks, where frequently flooded packets have only a limited impact on network performance. Controlling the flooding area can help to refine the technique. This can be achieved by gradually increasing, until the node is located, the number of hops involved in the flooding propagation.

5.2. Ad hoc routing and forwarding

The highly dynamic nature of a mobile ad hocnetwork results in frequent and unpredictable changes of network topology, adding difficulty and complexity to routing among the mobile nodes. The challenges and complexities, coupled with the critical importance of routing protocol in establishing communications among mobile nodes, make routing area the most active research area within the MANET domain. Schemes have been proposed to alleviate this problem by reducing redundant broadcasting. surveys existing methods for flooding a wirelessnetworkintelligently.Unicastforwarding means a one-to-one communication,i.e., one source transmits data packets to a single destination. This is the largest class of routing protocols found in ad hoc networks.

6. Applications and middleware

MANET flexibility makes this technology attractive for several applicative scenarios like, for example, in personal area networking, home networking, law enforcement operation, search and-rescue operations, commercial and educational applications, sensor networks . Table 2provides a categorization of present and possible future applicative scenarios for MANETs, as well as the services they may provide in each area.

6.1. Middleware

The middleware layer operates between the networking layers and the distributed applications. This constitutes major complexity inefficiency in the development of MANET applications. Research on middleware for mobile ad hoc networks is still in its infancy. Ad hoc networking and self-organization have not yet received the attention they deserve. Existing middleware mainly focus on mobile/nomadic environments, where a fixed infrastructure contains the relevant information.

7. Cross layers' research issues

These include among others energy conservation, security and cooperation, simulation and performance evaluation, and QoS.

7.1. Energy conservation

Power-saving policies at the operating system level include strategies for CPU scheduling, and for the hard-disk management .At the application-level, policies that exploit the application semantic or profit of tasks remote execution have been proposed. However, in small mobile devices, networking activities have a major impact on energy consumption.

7.2.2. Security at data link layer:

Security in the IEEE 802.11 standard is provided by the Wired Equivalent Privacy (WEP)scheme. WEP supports both data encryption and integrity. The security is based on a 40-bit secret key. Since WEP does not provide any support for the exchange of pair-wise secret keys, the secret key must be manually installed on each device. Asleep suffers from various design flaws and weaknesses, to correct the WEP problems a task group part of the IEEE 802.11i standardization is designing the new 802.11 security architecture.

8.Future Directions:

This paper focuses on the performance metric of*throughput*without taking into consideration *delay*. The delay experienced by the packets under the strategy proposed in this paper is large, increasing with the size of the system. As such, the result should be viewed as a theoretical one. What the theory does suggest is that for delay tolerant applications, there is ample opportunity to trade off delay and throughput by exploiting mobility. The result of this paper can be considered as an extreme point in the tradeoff, without any constraint on the delay. With a tighter delay constraint, the maximum achievable throughput must decrease. It would be interesting to characterize the optimal tradeoff between throughput and delay and to determine the kind of strategies that achieves this tradeoff. The ideas in this paper are not very relevant to real-time applications such as voice communications. However, wireless data services are expected to grow quickly over the next few years. A subset of these services, such as email and database synchronization, does indeed possess very loose delay constraints

(on the order of hours). Also, wireless devices are bound to become smaller and more pervasive in the future; they will not only be

carried by humans, but integrated into physical objects (such as cars, electrical appliances, etc.) It is unlikely that the density of base-stations will keep pace, due to regulatory and environmental hurdles in deploying them. Thus, there is a clear opportunity for wireless ad hoc networks to extend the reach of wireless communication. Our results suggest that delay-tolerant applications can take advantage of node mobility to significantly increase the throughput capacity of such networks.

9. Conclusions

MANET WG proposes a view of mobile ad hocnetworks as an evolution of the Internet. This mainly implies an IP-centric view of the network, and the use of a layered architecture. Current research points out though that this choice may limit developing efficient solutions for MANET. Other promising directions have been identified .The use of the IP protocol has two main advantages: it simplifies MANET interconnection to the Internet, and guarantees the independence from wireless technologies. The layered paradigm has highly simplified Internet design, however when applied toad hoc networks, it may result in poor performances it prevents exploiting important interlayer dependencies in designing efficient ad hoc network functions. For example, from the energy management standpoint, power control and multiple antennas at the link layer are coupled with power control and scheduling at MAC layer, and with energy-constrained and delay-constrained routing at network layer . Relaxing the Internet layered architecture, by removing the strict layer boundaries, is an open issue in the MANET evolution. Cross-layer design of MANET architecture and protocols is a promising direction for meeting the emerging application requirements, particularly when energy is a limited resource. From the economic standpoint, the main question to be addressed in the MANET model is the identification of business scenarios that can move MANET_s success beyond the academy and labs. potentialities cannot become a reality without an economic model that identifies potential revenues behind MANET-based network services.

With the current 802.11 technology, higher speeds imply a reduction in the coverage area of the Access Point (AP). Spreading in a hotspot a large number of APs to guarantee the coverage is not appealing both from the economic (infrastructure cost) and technical standpoint (APsinterference). The ad hoc paradigm can possibly offer an efficient solution to this problem: the APsupgraded with multi-rate high-speed technologies (e.g., 802.11a) achieve the required coverage by exploiting a multi-hop wireless network. While from a technology standpoint, feasible solutions can be designed to apply the MANET technology to extend APs_ coverage; the critical point remains the economic model. Which model could be applied for example in such a scenario to have users cooperating to provide support to the network service provisioning remains a question that typifies the open issues on the way of transitioning MANET results into the business environment.

References

[1] J. Ahola, Ambient Intelligence, ERCIM (European Research Consortium for Information and Mathematics) NEWS, N. 47, October 2001.

[2] A. Ahuja et al., Performance of TCP over different routing protocols in mobile ad-hoc networks, in: Proceedings of IEEE Vehicular Technology Conference (VTC2000), Tokyo, Japan, May 2000.

[3] G. Anastasi, E. Borgia, M. Conti, E. Gregori, IEEE802.11 ad hoc networks: performance measurements, in: Proceedings of the Workshop on Mobile and Wireless Networks (MWN 2003) in conjunction with ICDCS 2003,May 19, 2003.

[4] G. Anastasi, M. Conti, E. Gregori, IEEE 802.11 ad hocnetworks: protocols, performance and open issues, in: S.Basagni, M. Conti, S. Giordano, I. Stojmenovic (Eds.),Ad hoc Networking, IEEE Press Wiley, New York,2003.

[5] G. Anastasi, M. Conti, E. Gregori, A. Passarella, A power saving architecture for web access from mobile computers, in: Proceedings of the Networking 2002, Lecture Notes in Computer Science, vol. 2345, Springer, Berlin, 2002,pp. 240–251.I. Chlamtac et al. / Ad Hoc Networks 1 (2003) 13–64 53

[6] S. Basagni, Distributed and mobility-adaptive clustering for multimedia support in multi-hop wireless networks, in: Proceedings of the IEEE Vehicular Technology Conference (VTC) 1999, Amsterdam, The Netherlands, September19–22, 1999.

[7] C. Bettstetter, Smooth is better than sharp: a random mobility model for simulation of wireless networks, in: Proceedings of the ACM Workshop on Modeling, Analysisand

Simulation of Wireless and Mobile Systems,2001, pp. 19–27.

[8] C. Bisdikian, An overview of the Bluetooth wireless technology, IEEE Communication Magazine, December2001.

[9] A.V. Bakre, B.R. Badrinath, Implementation and performance evaluation of indirect TCP, IEEE Transactions on Computers 46 (1997).

[10] L. Blazevic, L. Buttyan, S. Capkun, S. Giordano, J.-P.Hubaux, J.-Y. Le Boudec, Self-organization in mobile adhoc networks: the approach of terminuses, IEEE Communication Magazine, June 2001, pp. 166–175.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

OPTIMAL CONTENT DOWNLOADING IN VEHICULAR NETWORKS

[1] C.P Balasubramaniam M.Sc(CS).,M.Phil., [2] T.Priya M.Sc(IT).,

[1]Assistant Professor [2]Research Scholar,
[1 & 2] Kongu Arts And Science College, Erode.

ABSTRACT

We consider a system where users aboard communication-enabled vehicles are interested in downloading different contents from Internet-based servers. This scenario captures many of the infotainment services that vehicular communication is envisioned to enable, including news reporting, navigation maps, and software updating, or multimedia file downloading. In this paper, we outline the performance limits of such a vehicular content downloading system by modeling the downloading process as an optimization problem, and maximizing the overall system throughput. Our approach allows us to investigate the impact of different factors, such as the roadside infrastructure deployment, the vehicle-to-vehicle relaying, and the penetration rate of the communication technology, even in presence of large instances of the problem. Results highlight the existence of two operational regimes at different penetration rates and the importance of an efficient, yet 2-hop constrained, vehicle-to-vehicle relaying.

Index Terms—Vehicular networks, downloading, optimization

1 INTRODUCTION

The presence of high-end Internet-connected navigation and infotainment systems is becoming a reality that will easily lead to a dramatic growth in bandwidth demand by invehicle mobile users. Examples of applications of vehicular communication abound, and range from the updating of road maps to the retrieval of nearby points of interest, from the instant learning of traffic conditions to the download of touristic information and media-rich data files. This will induce vehicular users to resort to resource-intensive applications, to the same extent as today's cellular customers. Most observers concur that neither the current nor the upcoming cellular technologies will suffice in the face of suchsurge in the utilization of resource-demanding applications. Recent network overload episodes incurred in by cellular infrastructures in presence of smartphone users provide a sobering wake-up call. To wit, a recent analysis on the traffic of a large US-based operator showed that smartphone users represent just 1 percent of the total subscribers, yet drain 60 percent of the network resourcesTo design a network architecture that will scale to support the mass of vehicular users, one possibility is to offload part of the traffic to Dedicated Short-Range Communication (DSRC), through direct infrastructure-to-vehicle (I2V) transfer, as well as vehicle-to-vehicle (V2V) data relaying. Such an approach is especially attractive in the case of the download of large amounts of delay-tolerant data, a task that is likely tochoke 3G/4G operator networks, but that well fits DSRCbased I2V and V2V communication paradigms due to its lack of strict time constraints.

2. RELATED WORK

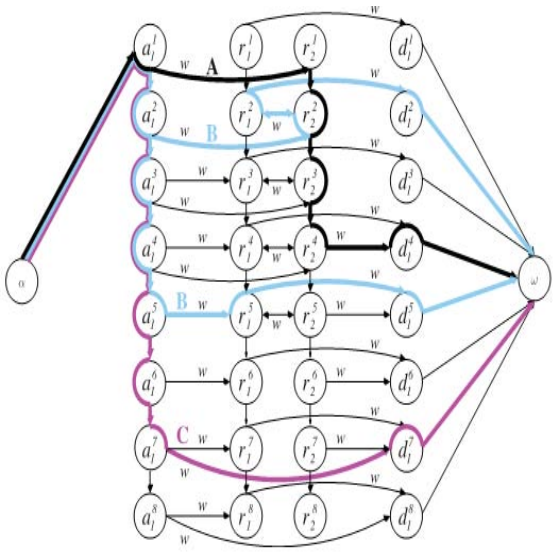
Infrastructure deployment. Earlier studies [9], [10] focus on the feasibility of using IEEE 802.11 APs to inject data into vehicular networks, as well as on the connectivity challenges posed by such an environment. he objective is not content downloading but the dissemination of information to vehicles in the shortest possible time. In infrastructure placement strategies are proposed that maximize the amount of time a vehicle is within radio range of an AP. Although longer periods of time under coverage can undoubtedly favor the download of contents by vehicular users, important differences with our work exist. First, our analysis is not limited to direct transfers from APs to vehicles, but includes traffic relaying. Second, while the problem formulation in [3] guarantees a minimum coverage requirement and the one in [4] maximizes the minimum-contact opportunity, we optimize the actual throughput.

Content downloading and dissemination.focuses on the access to web search and presents a system that makes such a service highly efficient by exploiting prefetching. Experimental and analytical results show the contribution of V2V and I2V communications to the system performance.

Our objective is to design the content downloading system so as to maximize the aggregate throughput. To this aim, we have to jointly solve two problems: 1) given a set of candidate locations and a number of APs to be activated, we need to identify the deployment yielding the maximum throughput; 2) given the availability of different data transfer paradigms, possibly involving relays, we have to determine how to use them to maximize the data flow from the infrastructure to the down loaders. Our approach consists in processing a road layout and an associated vehicular mobility trace, so as to build a graph that represents the temporal network evolution (Section 4). By using this graph, we formulate a max-flow problem whose solution matches our goals.

4. DYNAMIC NETWORK TOPOLOGY GRAPH (DNTG)

We generate a time-expanded graph [26], hereinafter DNTG, from a vehicular mobility trace. To build the graph, we consider that on the road layout corresponding to the mobility trace there are: 1) a set of A candidate locations (a;i ¼ 1; ...;A) where APs could be placed, 2) a set of V vehicles (v I ;i ¼ 1; ...;V) transiting over the road layout and participating in the network, and 3) a subset of D vehicles that wish to download data from the infrastructure.



(b) DNTG resulting from the contact events

Each contact event is characterized by:

1. The quality level of the link between the two nodes. Several metrics could be considered; here, we specifically take as link quality metric the data rate achievable at the network layer;

2. the contact starting time, i.e., the time instant at which the link between the two nodes is established or the quality level of an already established link takes on a new value;

3. the contact ending time, i.e., the time instant at which the link is removed, or its quality level has changed.

We stress that, by associating a time duration to the contact events, instead of considering them as atomic, we can model critical aspects such as channel contention.

5. THE MAX-FLOW PROBLEM

Given the DNTG, our next step is the formulation of an optimization problem whose goal is to maximize the flow from to !, i.e., the total amount of downloaded data.

$$\max \sum_{k=0}^{n} \mathbf{X(V, w)}$$

Denoting by xð; Þ, the traffic flow over an edge connecting two generic vertices,

5.1 Constraints

Nonnegative flow. The flow on every existing edge must be greater than or equal to zero.

Flow conservation. For any vertex in the DNTG, the amount of incoming flow must equal the amount of outgoing flow. This constraint is expressed in slightly different form, depending on whether the vertex represents a downloader, a relay, or a candidate AP location.

Channel access. As mentioned, we deal with unicast

transmissions and assume that the nodes use a 802.11-based MAC scheme; also, V2V and I2V communications occur on different channels. Then, given a tagged vehicle, we consider that none of the following events can take place simultaneously, and the time span of each frame must be shared among them:

1. The vehicle transmits to a neighboring vehicle, 2. a neighboring vehicle receives from any relay, 3. the vehicle receives from a neighboring relay, 4. a neighboring relay transmits to any vehicle, and 5. the vehicle receives from a neighboring AP.

Recall that we do not model the scheduling of the single packets transmitted within each frame.

Overlapping AP coverages. Recall that, when a vehicle falls

within coverage of two or more APs, we assume that, during a frame, it communicates with one AP only, and that the APs operate on different frequency channels.

We leverage the capability of our framework to model different AP deployments and explore the following strategies:

Random: A locations are randomly selected among the candidate ones, according to a uniform distribution.

Crowded: A locations are picked, whose coverage area exhibits, over time, the highest vehicular density.

Contact: A locations are selected, which maximize the sum of the contact opportunities between vehicles and APs. Inspired by the metric adopted in [4], for each vehicle we express the contact opportunity as the fraction of the road segment lengths traveled while under coverage of at least one AP.

Throughput. The average per-downloader throughput,, is very satisfying in all conditions, scoring well above 10 Mb/s even in low-p, low- A scenarios, and more than 20 Mb/s in presence of a wide I2V and V2V technology adoption.

Delay. Delays, in Fig. 5b, are in the order of tens of seconds, a good result when considering the delay-tolerant nature of most V2V transfers. We can observe different behaviors for low and high values of p. For p<20%,an increase in availability of relays leads to more frequent V2V transfers, hence higher delays. When p>30%, the already pervasive presence of relays makes the impact of even higher

penetration rates negligible. Also, a denser AP deployment helps reducing the delay, although such a gain is significant only when A is low.

Fairness. To get an insight on how the system throughput is actually shared among the downloaders, shows the Jain's fairness index. The increase in penetration rate has a major impact because it implies a growing number of V2V communication opportunities.

Transfer paradigm. The above observations on the funda-mental role of V2V traffic relaying is confirmed by the results in Fig. 5d, depicting the fraction of content downloaded through relay vehicles. Indeed, most of the content is received by downloaders from relays (through either connected forwarding or carry-and-forward).

Summary. The performance metrics are consistent in revealing the critical importance of the penetration rate p and the lower impact of the roadside infrastructure extension. Accordingly, we can separate two regimes. The first, when p<20%, i.e., at early stages of the technology adoption, characterized by lower throughput and higher delay, a stronger dependency on direct I2V communication, and lower downloading fairness. The second, for p>30%, i.e., in presence of a quite mature technology, featuring instead higher throughput and lower delay, massive use of V2V communication and high fairness.

7.1 Impact of Vehicular Communication Technology Adoption

As a first step in the evaluation of vehicular content downloading, we look at the impact that the diffusion of I2V and V2V communication technologies has on the system performance. To that end, we consider different values of p as well as different extensions of the roadside

7.2 Low-Penetration Regime

As case study of the low-penetration regime, we consider p ¼ 10%. The default settings include the urban scenario, non overlapping AP coverages, a 1 percent fraction of downloaders and a 2-hop limit in V2V relaying. AP deployment average per-downloader throughput and delay for the different deployment strategies, as the number of active APs A varies. Overall, the performance at early deployment stages is satisfactory. The plots confirm that increasing A positively affects the downloading performance.

Unlimited relaying. All previous results assumed a 2-hop limit in data transfers, basically constraining V2V relayingone hop at most. We now relax this assumption and compare three different scenarios, where 1) only direct I2V communication is allowed, 2) the 2-hop limit is enforced, and 3) unlimited relaying is allowed.

Road environment. The average per-downloader throughput recorded in the three road topologies presented in Section 6 is portrayed in Fig. 11a, when the APs are deployed as dictated by the Max-flow strategy. The overall performance trend in the new environments is the same as already observed in the urban scenario; thus, our considerations on the impact of the AP deployment also hold for the village and suburban environments.

However, we can observe that the relative result in suburban and village environments differs from that measured in the urban case. On the one hand, the throughput in the village scenario is lower than in the urban one, with a significantly reduced utilization of V2V relaying. On the other, the suburban scenario yields higher V2V download fraction and per-downloader throughput. The reason for these different behaviors is found in the diverse nature of vehicular traffic in the tree regions. By looking at Fig. 4b, it is clear that fewer vehicles circulate in the village environment than in

the urban one: Thus, for a given p, fewer vehicles participate in the content downloading as relays. Moreover, the traffic is distributed over the road topology quite evenly, which makes it difficult to find an AP deployment that well covers most of the vehicular traffic. As a result, downloaders in the village scenario are penalized in terms of throughput.

Concurrent downloaders. In presence of a wide diffusion of I2V and V2V communications, the downloading activity by users participating in the system is likely to grow. Thus, in the high-penetration regime, it is important to evaluate the impact of the amount of concurrent downloaders, i.e., users requesting some content during the same time interval.

Road environment. the throughput and V2V download fraction in the urban, suburban, and village scenarios.

CONCLUSION

We proposed a framework based on time-expanded graphs for the study of content downloading in vehicular networks. Our approach allows to capture the space and time network dynamics, and to formulate a max-flow problem whose solution provides an upper bound to the system performance**.**

REFERENCES

[1] B.S. Arnaud, "iPhone Slowing Down the Internet - Desperate Need for 5G R&E Networks," http://billstarnaud.blogspot.com/2010/ 04/iphone-slowing-down-internet- esperate.html, May 2010.

[2] U. Paul, A.P. Subramanian, M.M. Buddhikot, and S.R. Das, "Understanding Traffic Dynamics in Cellular Data Networks," Proc. IEEE INFOCOM, Apr. 2011.

[3] Z. Zheng, P. Sinha, and S. Kumar, "Alpha Coverage: Bounding the Interconnection Gap for Vehicular Internet Access," Proc. IEEE INFOCOM, Apr. 2009.

[4] Z. Zheng, Z. Lu, P. Sinha, and S. Kumar, "Maximizing the Contact Opportunity for Vehicular Internet Access," Proc. IEEE INFOCOM, Mar. 2010.

[5] F. Malandrino, C. Casetti, C.-F. Chiasserini, and M. Fiore, "Content Downloading in Vehicular Networks: What Really Matters," Proc. IEEE INFOCOM Mini-Conf., Apr. 2011.

[6] G. Iosifidis, I. Koutsopoulos, and G. Smaragdakis, "The Impact of Storage Capacity on End-to-End Delay in Time Varying Networks," Proc. IEEE INFOCOM, Apr. 2011.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

An Efficient and Enhanced path sequence algorithm for missing pages in Community Web

[1]R.Kousalya **[2] S.Mohanapriya**

[1] P.hd.Scholar/HOD,Department of Computer Applications,

Dr.N.G.P. Arts and Science College, Coimbatore,Tamilnadu

[2]Asst Professor/Department of BCA , PSGR Krishnammal College for Women, Coimbatore,Tamilnadu

ABSTRACT

Web directory is a concept it lists web sites by category and subcategory. The Web content into thematic hierarchies these are known as Web Directories and correspond to listings of topics which are organized and overseen by humans. Web personalization is a technique websites are personalized based on the interests of an individual, group or organization. In context previous work the Web directory is viewed as a thematic hierarchy and personalization on the basis of usage data. In dissimilarity to most of the work on Web usage mining, the usage data that are analyzed here communicate to user navigation throughout the Web. Due to proxy servers and cached versions of the pages used by the client using 'Back', the sessions identified have many missed pages. In this work Enhanced Path Sequence Algorithm proposed there are chances of missing pages after constructing transactions due to proxy servers and caching problems. Three approaches used for this 1. Time Window: 2. Reference Length approach: 3. Maximal Forward Reference.

Keywords: Web Mining, Personalization, Clustering, Data Cleaning, Log File

INTRODUCTION

The Web has not achieved its goal of providing easy access to online in sequence and its size is growing, the great quantity of available information on the Web causes the frustrating phenomenon of "information overload" to Web users. The information for the topic that a user is seeking might reside very deep within the directory. Therefore, the size and the complexity of the Web directory itself are canceling out the gains that were expected with respect to the information overload problem and it is often difficult to navigate to the information of interest to a particular user. On the other hand, Web Personalization the task of making Web-based information systems adaptive to the needs and interests of single users, or more of users, emerge as an important means to tackle information overload. Due to proxy servers and cached versions of the pages used by the client using 'Back', the sessions identified have many missed pages. However, in achieving personalization, are confronted with the difficult task of acquiring and creating accurate and prepared user models. Reliance on physical creation of the models, either by the user or by area experts, is inadequate for various reasons, among which the irritation of the users and the difficulty of verifying and maintaining the resulting models. Web Usage Mining is an approach that employs knowledge discovery from usage data to automate the creation of user models. In previous work explained the use of machine learning techniques for modeling the user communities based on clustering and probabilistic modeling. However, these personalized directories suffer from what we call the "local overload" problem.

Problem Formation

1. The size and the complexity of the Web directory itself are canceling out the gains that were expected with respect to the information overload problem
2. These personalized directories suffer from what we call the "local overload" problem.

To overcome the deficiencies of Web directories and Web personalization by combining their strength, as long as a new tool to battle information overload. In particular, this research focus on the construction of usable Web directories that models the interests of groups of users, known as user communities. In this work presents a knowledge discovery framework for constructing community-specific web directory. A new objective of Web personalization addressing this issue, introduce a novel methodology that combines usage data and

thematic information from web directory. These methods enhance the proposed methods, using an approximation of the amount of thematic information included a priori in the categories of the Web directory. This in order is exploited by the introduction of a new criterion in our methodology, most important to the construction of more fine grained community Web directories, and thus, to an efficient tackle of the limited overload problem. The enhancement of these methods, a new one is obtainable that combine clustering and probabilistic modeling by applying Probabilistic Latent Semantic Analysis (PLSA) to the results derived by a clustering algorithm. Path completion technique is which is used to append lost pages and construction of transactions in preprocessing stage Path completion process which is used to append lost pages and construction of transactions in preprocessing stage.

I. Methodology

1. Usage Data Preparation

The usage data that form the basis for the construction of community Web directories are collected in the access log files of ISP cache proxy servers and data record the direction-finding of the subscribers through the Web and they are usually diverse and voluminous. The outgoing traffic is higher than the usual incoming traffic of web site and the web pages more diverse semantically. The task of usage data preparation is to assemble these data into a steady, integrated, and short view.

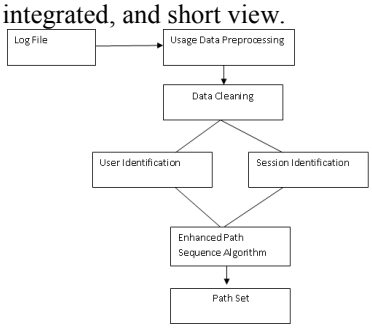


Fig 1: **Architecture Diagram**

After that stage is the identification of individual user sessions. The fact that are focusing on the discovery of behavioral patterns in the data, somewhat than individual users, allowed us to overcome the lack of user registration data or other means of user identification, such as cookies, and lead us to exploit a simple kind of user session. User session is define as a sequence of log entries accesses to web pages by the same IP address and where the time interval between two subsequent entries does not exceed a certain time threshold.

2. Data Cleaning

The process of data cleaning is removal of outliers or unrelated data. Analyzing the huge amounts of records in server logs is a unwieldy activity. So early cleaning is necessary. If a user requests an exact page from server entries like gif, JPEG images are downloaded which are not of use for further analysis are eliminate. The records with failed status code are also eliminated from logs. Automatic programs like net robots, spiders and crawlers are also to be removed from log files.

3. Enhanced Path sequence algorithm

Due to proxy servers and cached versions of the pages used by the client using 'Back', the sessions identified have many missed pages. So path completion step is carried out to identify missing pages. Path Set is the incomplete accessed pages in a user session. It is extracted from every user session set. There are chances of missing pages after constructing transactions due to proxy servers and caching problems. Three approaches used for this 1. Time Window: A time window transaction is framed from triplets of ip address, user identification, and time length of each webpage up to a limit called time window. 2. Reference Length approach: This approach is based on the assumption that the amount of time a user spends on a page correlates to whether the page is an auxiliary page or content page for that user. 3. Maximal Forward Reference: A transaction is considered as the set of pages from the visited page until there is a backward reference. Forward reference pages are considered as content pages and the path is taken as index pages. To solve this an Enhanced Path Sequence Algorithm is proposed in this work to generate accurate path sequences by using two way hashed structure based access history list to frame a complete path with optimal time. In this Session Identification algorithm data structures such as Array List to represent Web Logs and User Access List, a Hash table to represent server pages, a two-way hashed structure are utilized. Two way hashed structure is used to store Access History List (AHL) to represent user accessed page sequence. Two hash tables primary and secondary hash tables are used in which primary is used to store sessions and pointers to secondary table which is having a complete path navigation. Access path set (AP) every user id (UserId) identified from user session (USession). It is defined as

AP = (UserId,(URI ,Date , RefeLength),… (URI(n), Date(n), RefeLength (n))

Where Refelength is computed for every record in data cleaning stage. After identifying path for each UserId path combination is done if two consecutive pages are same. In the user session if any of the URL specified in the Referrer URL is not equal to the URL in the previous record then that URL in the Referrer Url field of current record is inserted into this session and thus path sequence is obtained. The next step is to determine the reference length of new appended pages during path sequence and modify the reference length of adjacent ones. Since the assumed pages are normally considered as auxiliary pages the length is determined by the average reference length of auxiliary pages. The reference length of adjacent pages is also adjusted.

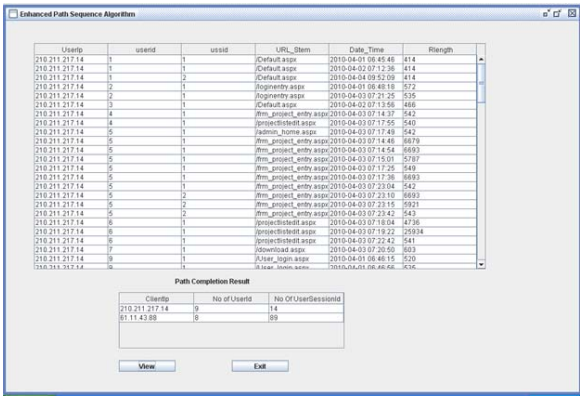


Fig 3: Path Sequence Algorithm

4. Community Directory Miner

Objective-Community Directory Miner is based on the cluster mining algorithm which has been employed earlier for site-specific community discovery. Cluster mining algorithm, named Community Directory Miner discovers patterns of common behavior by looking for all maximal fully connected sub graphs (cliques) of a graph that represents the user characteristic features. In first step it constructing the graph, the vertices it correspond to the categories, category co-occurrence in thematic session sets. Vertices and edges are connected with weights computed based on the category occurrence and co-occurrence frequencies. The connectivity of the graph is usually high. For this reason make use of a connectivity threshold that reduces the edges of the graph. The connectivity of the graph has been reduced the weighted graph is turned to an un-weighted one. At last all maximal cliques of the un-weighted graph are generated and each one corresponding to a community model. The main advantage of this approach is that each user may be assigned to many communities. The clusters generated by Objective-Community Directory Miner group together characteristic features of the user. Each clique discovered by Objective-Community Directory Miner is thus already a community model that is a set of interesting categories.

5. Clustering And Probabilistic Directory Miner

The enhanced methods presented above this module introduce here a new hybrid method for the discovery of community models. This technique combines a clustering algorithm with PLSA. Here the popular k-means clustering algorithm applied on the relation for the creation of the initial communities. It differs from Community Directory Miner clustering, as it produces non-overlapping clusters, i.e., each category belongs to only one cluster. The PLSA model is based on the assumption that there exist set latent factors such that each instances each observation of a certain category inside a thematic session tree is related to a latent factor. The Expectation-Maximization (EM) algorithm to estimate the required probabilities which correspond to the probability of a latent factor, the probability of a thematic session set, given the latent factor, and the probability of accessing Web pages of a certain category given the latent factor respectively. There is some hidden associations exist in the data, i.e., sub-communities inside the cluster that are not directly observable. To discover this hidden knowledge, map each cluster derived by k-means onto a new space of latent factors. The community Web directories are constructed using both observable and latent associations in the data and allow us to better model the interests of users.

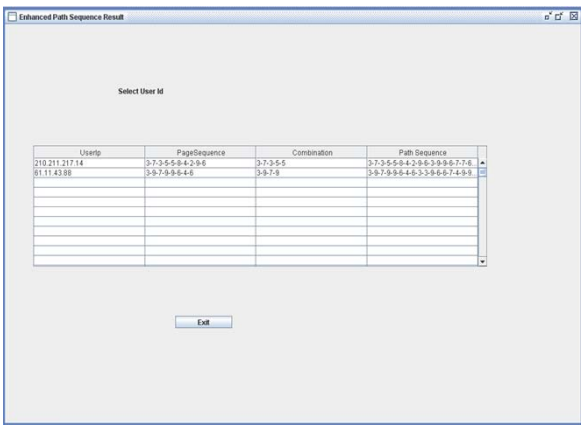


Fig 4: Enhanced Path Sequence Result

Conclusion:

In this paper proposed the concept of a community Web directory, as that specializes to the needs and interests of particular user communities. It presents the complete methodology for the construction of such directories with the aid of machine learning methods. A data preprocessing treatment system for web usage mining has been analyzed and implemented for record data. It has undergone various steps such as data cleaning, user identification, session discovery, and path completion and transaction detection. User community models take the form of thematic hierarchies and are constructed by employing clustering and probabilistic learning approaches. Path completion technique is added to solve the missing page problem Path Set is the incomplete accessed pages in a user session. It is extracted from every user session set.

References

[1] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," Comm. ACM, vol. 43, no. 8, pp. 142-151, 2000.

[2] J. Srivastava, R. Cooley, M. Deshpande, and P.T. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations, vol. 1, no. 2, pp. 12-23, 2000.

[3] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos, "Web Usage Mining as a Tool for Personalization: A Survey," User Modeling and User-Adapted Interaction, vol. 13, no. 4, pp. 311-372, 2003.

[4] G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C.D. Spyropoulos, "Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques," Interacting with Computers J., vol. 14, no. 6, pp. 761-791, 2002.

[5] G. Xu, Y. Zhang, and Y. Xun, "Modeling User Behaviour for Web Recommendation Using lda Model," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence and Intelligent Agent Technology, pp. 529-532, 2008.

[6] W. Chu and S.-T.P. Park, "Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models," Proc. 18th Int'l Conf. World Wide Web (WWW), pp. 691-700, 2009.

[7] The Adaptive Web, Methods and Strategies of Web Personalization, P. Brusilovsky, A. Kobsa, and W. Neijdl, eds. Springer, 2007.

[8] D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis, and M. Dikaiakos, "Web Community Directories: A New Approach to Web Personalization," Web Mining: From Web to Semantic Web, B. Berendt et al., eds., pp. 113-129, Springer, 2004.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

ANALYSIS OF ELLIPTIC CURVE INTEGRATED ENCRYPTION SYSTEM IN VANET

[(1)] C.Kumuthini MCA., M.Phil., [(2)]K.Tharani, , [(3)] P.Krishnakumari

[1]Phd Scholar, [2] Research Scholar, [3] Director of MCA
[1 & 2] Dr. N.G.P. Arts and Science College.
.[3]RVS college of Arts and Science

ABSTRACT

It refers to a network connection established for a single session and does not require a router or a wireless base station. A VANET (Vehicular Adhoc NETwork) is a sub- class of mobile network. It is used to communicate between vehicles and road side units. Every car it acts as a wireless router or node it can be used to create a network within a range of 100 to 300 metres. If the car signal or dropping a network the other remaining cars or vehicles can join to form with the help of adhoc network to create a mobile internet but it does not rely on pre existing infrastructure. The main aim of the technology is used to provide to communicate between one vehicles to another vehicle to protect the public safety of the society.To provide secure over the internet, the several types of encryption and decryption techniques are used. In this paper, discuss about the several types of security algorithms in vanet and its possible solutions by using ECIES using encryption and decryption of the cryptographic scheme.

Keywords: Encryption,Decryption, Authentication, Vanet, SSL &SSH, RSA, ECC ECIES.

I. INTRODUCTION

Vehicular Ad hoc Networks is a special kind of mobile ad hoc network to provide communication among nearby vehicles and between vehicles equipments. It is mainly used for improving efficiency of vehicles and safety of (future) vehicles. There are number of possible attacks in VANET due to open nature of wireless medium. This security attacks and logically organized/represented in a more lucid manner based on the level of effect of a particular security attack on intelligent vehicular traffic.

Network security

It consists of the provisions and policies adopted by a network administrator to prevent and monitor unauthorized access, misuse, modification, or denial of a computer network and network-accessible resources. Network security involves the authorization of access to data in a network, which is controlled by the network administrator. Users choose or are assigned an ID and password or other authenticating information that allows them access to information and programs within their authority. Network security covers a variety of computer networks, both public and private, that are used in everyday jobs conducting transactions and communications among businesses, government agencies and individuals.

Networks can be private, such as within a company, and others which might be open to public access. Network security is involved in organizations, enterprises, and other types of institutions. It secures the network, as well as protecting and overseeing operations being done. The most common and simple way of protecting a network resource is by assigning it a unique name and a corresponding password[1].

II. PERFORMANCE COMPARISON OF ECC AND RSA ENCRYPTION SCHEMES

Rob Shaw and Yin (1997) compared the operational characteristics of RSA and ECC. In their article, they claimed that assuming all necessary parameters and with ECC will be almost 8 times longer than with RSA, and decryption will be almost 6 to 7 times faster[2]. These findings, however, contrasted that of Certicom Corporation who adjudged the most efficient implementation of ECC 10 times faster than comparable RSA systems[3].

2.1Theory of RSA Cryptosystem:

The RSA is the most widely used cryptosystem today. Unfortunately, encrypting a message, m, involves exponentiation, $\mathbf{C= m^e mod\ n}$, a mathematical procedure which requires a lot of

computations, making it impossible to achieve the speeds of private key systems such as DES, a phenomenon that is true for all public key systems[4].

To set up a RSA cryptosystem, a user (say Alice) picks two large primes p and q and computes their product, **n=pq** . The group used is the multiplicative group (g=z*) of units in the integer modulo n. It is well known that the order of G is ***φ=(p-1)(q-1)***, where φ denotes the Euler phi function. Clearly, Alice's public key is the pair of integers {n, e} and her private key is d.

2.2RSA Key Generation

An RSA key pair can be generated using key pair generation Algorithm. The public key consists of a pair of integers (n, e) where the RSA modulus n is a product of two randomly generated (and secret) primes p and q of the same bit length. The encryption exponent e is an integer satisfying $1 < e < \varphi$ and gcd $(e, \varphi) = 1$ where $\varphi = (p-1)(q-1)$. The private key d, also called the decryption exponent, is the integer satisfying $1 < d < \varphi$ and $ed \equiv 1 \pmod{\varphi}$.

It has been proven that the problem of determining the private key d from the public key (n, e) is computationally equivalent to the problem of determining the factors of p and q.

2.3 RSA key pair generation

INPUT: Security parameter l.
OUTPUT: RSA public key (n, e) and private key d.

1. Randomly select two primes p and q of the same bit lengthl/2.
2. Compute $n = pq$ and $\varphi = (p-1)(q-1)$.
3. Select an arbitrary integer e with $1 < e < \varphi$ and gcd $(e, \varphi) = 1$.
4. Compute the integer d satisfying $1 < d < \varphi$ and $ed \equiv 1 \pmod{\varphi}$.
5. Return (n, e, d).

2.4 RSA Encryption/ Decryption Scheme

RSA encryption schemes use the fact that **$med \equiv m \pmod{n}$**for all integers m. The encryption and decryption procedures for the (basic) RSA public-key encryption scheme are presented as basic RSA encryption and decryption Algorithms.

Decryption works because $cd \equiv (me)d \equiv m \pmod{n}$, as derived from expression. The security relies on the difficulty of computing the plaintext m from the cipher textc = me mod n and the public parameters n and e. This is the problem of finding e-th roots modulo n and is assumed (but has not been proven) to be as difficult as the integer factorization problem[6].

Basic RSA encryption:

INPUT: RSA public key *(n, e)*, plaintext $m \in [0,-1]$.
OUTPUT: Cipher text*c*.

1. Compute $c = me \bmod n$.
2. Return(*c*).

Basic RSA decryption:

INPUT: RSA public key *(n, e)*, RSA private key *d*, cipher text*c*.
OUTPUT: Plaintext *m*.

1. Compute $m = c^d \bmod n$.
2. Return (*m*).

2.5 COMPARISON BETWEEN ECC AND RSA

To test and compare the performance characteristics of the RSA and ECC encryption algorithms, we independently tested each of the following three main components for timings: key generation, encryption and decryption. Timings are not absolute so each operation for every test parameter was run 20 times in order to reach satisfactory level of confidence interval.

A 99.9% confidence interval was calculated from the test results using the student T-distribution. We also measured the size of the data files used to store the encrypted results.

The parameters of the operations are:

a. the size of the applied key.
b. the size and content of the input data.

Tests were performed on Intel Pentium dual core 1.6GHZ machine with 512MB of RAM. The message used for encryption is the 100 byte text." ECDLP is believed to be harder than both the Integer Factorization and Discrete Logarithm Problems".

Estimates are given for parameter sizes providing comparable levels of security for RSA and EC systems. The parameter sizes, also called key sizes, that provide equivalent security levels for RSA and EC systems are as listed in Table 1.

Table 1: Comparable key sizes between ECC and RSA

ECC	RSA
160	**1024**
224	**2048**
256	**3072**
384	**7680**
512	**15360**

III. Elliptic Curve Integrated Encryption System (ECIES)

Integrated Encryption Scheme (IES) is a hybrid encryption scheme which provides semantic security against an adversary who is allowed to use chosen-plaintext and chosen-cipher text attacks. The security of the scheme is based on the Diffie–Hellman problem.The elliptic curve integrated encryption system (ECIES) is the standard elliptic curve based on encryption algorithm. It is called integrated, since it is a hybrid scheme that uses a public key system to transport a session key for use by a symmetric cipher. ECIES is a public-key encryption algorithm.

To send an encrypted message to Bob using ECIES Alice needs the following information:

- cryptographic suite to be used:
  - KDF
  - MAC
  - symmetric encryption scheme E
- EC domain parameters(p,a,b,G,n,h) for a curve over prime field or (m,f(x),a,b,G,n,h) for a curve over binary field;
- Bob's public key: KB(Bob generates it as follows: KB = KBG , where KB is the private key he chooses at random: KB€ [1, n-1]
- optional shared information: S1 and S2.

To encrypt a message m Alice does the following:

1. It generates a random number r € [1,n-1] and calculates R = rG;
2. we can derives a shared secret: S = Px, where P = (Px, Py) = rKB (and P ≠ 0)
3. It uses KDF to derive a symmetric encryption and a MAC keys: KE || KM = KDF(S||S1);
4. It can be encrypts the message: c = E(KE; m);
5. It can computes the tag of encrypted message and S2: d = MAC(KM; c||S2);

outputs R||c||d.

To decrypt the cipher text R||c||d Bob does the following:

1.It can be derives the shared secret: S = Px , where P = (Px, Py) = KBR (it is the same asthe one Alice derived because P = KBR = KBrG = rKBG =rKB), or outputs *failed* if P = 0;

2. We can derives keys the same way as Alice did: KE||KM = KDF (S||S1);

3.It uses MAC to check the tag and outputs *failed* if d ≠ MAC(KM ; c||S2);

4. It uses symmetric encryption scheme to decrypt the message m = E -1(KE; c)

IV. Proposed System

A process of converting Plain Text into Cipher Text is called as Encryption. A reverse process of encryption is called as Decryption.

To overcome the problems in Elliptic Curve cryptography is to provide the requirements:

a) Security is difficult
b)Size of elliptical curve cryptography
c)Discreate log the problem
d)Intractability of certain mathematical problem

In existing several algorithms or techniques are used for encryption and decryption it will be produce better result. In proposed system will be implemented by using ECIES with C++ it provides more security than previous one encryption and decryption methods.

V.Conclusion and Future Work

Public-key encryption can be used to eliminate problems involved with conventional encryption. However, it has not managed to be as widely accepted as conventional encryption because it introduces a lot of overheads. Therefore, it is very important to find ways to reduce the overheads yet not sacrificing on other aspects of security so that the desirability in public-key can be exploited.

After comparing the RSA and ECC ciphers, the ECC has proved to involve much less overheads compared to RSA. The ECC has been shown to have many advantages due to its ability to provide the same level of security as RSA yet using shorter keys. However, its disadvantage which may even hide its attractiveness is its lack of maturity, as mathematicians believed that enough research that has been compare to ECC is better than ECIES for encryption and decryption for security.

There are several techniques and algorithms implemented in MANET and as well as in VANET too they are: RSA algorithm, diffie-hellman Algorithm and Elliptic Curve Integrated Encryption System. By using Elliptic Curve Integrated Encryption System it can provide security at minimum level.

In future it can be implemented by using ECIES with C++ with the help of NS2 simulator tool or MAT LAB.

VI. REFERENCE

[1]http://en.wikipedia.org/wiki/Network security[2]Rob shaw, M. J. B. and Y. L. Yin (1997),“Elliptic CurveCryptosystems’, Http://www.rsasecurity.com/rsalabs/ecc/elliptic curve.html.

[3]Stallings, W. (2003)," Cryptography and Network Security: Principles and Practice", 3rd edition, PrenticeHall, New Jersey.
[4]Brown, M., D. L. Hankerson, J.L_opez and A. Menezes (2001), "Software implementation of the NIST Elliptic curves over prime fields.In Progress in Cryptology - CT-RSA, D. Naccache", Ed.,vol. 2020of Lecture Notes in Computer Science, pp. 250-265.
[5]Certicom Corp., (2004), "An elliptic curve cryptography (ecc)primer,White paper, erticom.
[6]Weil, N. (1998)," U.S. govt.'s encryption standard cracked in record time:NetworkWorld".http://www.networkworld.com/news0720des.html

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

ANALYSIS OF VARIOUS POSITION-BASED ROUTING PROTOCOLS AND GREEDY METHOD IN VANET

[1]C.Kumuthini [2]P.KrishnaKumari

[1]Research Scholar, Bharathiar University, Coimbatore
[2] Director , Department of M.C.A, R.V.S College of Arts and Science,Coimbatore

ABSTRACT

Vehicular Ad Hoc Network (VANET) is a sub class of Mobile Ad hoc Networks (MANET) , which provides wireless communication among vehicles and vehicle to road side equipments. Multi-hop data delivery between vehicles is an important aspect for the support of VANET-based applications. But the communication links in VANET are highly vulnerable to disconnection and hence the routing reliability of these ever-changing networks needs to be paid special attention. Current routing protocols usually use geographic information to forward packets as they perform well in reliable delivery. However, in a city network, intersections place a unique challenge to routing protocols. Thus some of the geographic routing protocols may include numerous intermediate intersections which results in routing paths with higher hop count. To avoid such problems hop greedy routing and back bone mechanism are used. The hop greedy routing protocol exploits the transmission range, avoids intersections that are used to change the direction of the routing path and ensures that the selected intersections have enough connectivity. In back-bone mechanism, the nodes perform functions such as detecting void regions on road segments, storing packets on unavailability of forwarding nodes and tracking the movement of end nodes. Thus this project focuses on designing an effective routing strategy for VANET in city environment.

Keywords

Vehicular adhoc network(VANET), Geographic routing, City Environment, Reliable delivery, Hop Greedy Routing.

1.INTRODUCTION

With the sharp increase of vehicles on roads in the recent years, driving becomes more challenging and dangerous. The main goal of Vehicular adhoc network (VANET) is providing safety and comfort for passengers and helping drivers on the roads by anticipating hazardous events or bad traffic areas. VANETs allow vehicles to avoid problems, either by taking any desired action or by alerting the driver. Several applications are provided for vehicle to vehicle communication (V2V) which exchange real time information to assist the drivers. A VANET turns every participating car into a wireless router or node, allowing cars approximately 100 to 300 metres of each other to connect and in turn, create a network with a wide range. As cars fall out of the signal range and drop out of the network, other cars can join in, connecting vehicles to one another so that a mobile Internet is created. Currently, intelligent transportation system components provide a wide range of services such as crash prevention and safety, driver assistance, freeway management, and infotainment for drivers and/or passengers[1].

Routing in a city network would not be exactly the same as in a highway. This is because routing in city environment has difficulties to handle two-dimensional scenarios with obstacles (buildings) and voids. Thus it is difficult to predict the exact traffic density of a region. In addition to this, intersections place a unique challenge for routing in city scenarios. The main concern is whether the performance of VANET routing protocols can satisfy the throughput and delay requirements of such applications.

2. NETWORK ARCHITECTURE AND CHARACTERISTICS

Wireless ad hoc networks do not depend on fixed infrastructure for communication and

dissemination of information. The architecture of VANET consists of three categories[2]:

- Inter-vehicle communication

  This is also known as pure adhoc networking. In this type of communication the vehicles communicate with each other with no infrastructure support.

- Vehicle-to-roadside communication

  This is also known as vehicle-to-infrastructure (V2I) communication. In this type of communication, the vehicles use cellular gateways and wireless local area network access points to connect to the Internet and to facilitate vehicular communication.

- Inter-roadside communication

  This is also known as hybrid communication. Vehicles can use infrastructure to communicate with each other and share the information received from infrastructure with other vehicles through ad hoc communication.

2.2 Specific Characteristics of VANETs

The unique features of VANETs when compared to MANETs are[10]:

- Higher transmission power and storage : In VANETs the nodes (vehicles) are usually equipped with higher power and storage than those in MANETs.
- Higher computational capability : Operating vehicles can afford higher computing, communication and sensing capabilities than MANETs.
- Predictable mobility : In VANETs the movement of the network nodes can be predicted as they move on a road network but this is not possible in MANETs. The future position of the vehicle can be predicted if the current velocity and road trajectory information are known.

2.1 Characteristics of City Scenarios

The main difference between highway and city scenarios is that highway scenario is largely one-dimensional which is very beneficial for greedy position-based routing approaches. Cities, however, face the following challenges to routing protocols[9]:

- **Geometric Two-Dimensionality:** In a city scenario, vehicles change their direction of movement frequently. Moreover, cars can move with varying relative angle allowed by the street geometry. In contrast to the highway, this weakens the correlation of the destination position to a suitable next-hop. **Obstacles:** A city is usually characterized by the presence of radio obstacles(buildings). This creates problems with position-based next-hop selection. Whenever the line-of-sight between two nodes touches an obstacle, the nodes are not able to communicate.
- **Node Density:** In cities of industrialized countries, the node density can be expected to be rather high with respect to the radio range, especially at “density hot spots” like junctions. On one hand, node density creates a better ad-hoc connectivity, but on the other hand, it poses a challenge to flooding mechanisms that need to be very efficient.
- **Low Mobility:** Compared to highway scenarios, the nodes move at rather low speeds in city roads, influenced by node density (the more nodes the slower the movement) and constrained by speed limits. Also, the mobility is location-dependent, e.g., it is lower at junctions.

3. RELATED WORK

To enhance the safety of drivers and provide the comfortable driving environment, messages for different purposes need to be sent to vehicles. Fast and reliable communication between cars (vehicle-to-vehicle) are essential for future vehicle alert systems. From a network perspective, this means that messages have to be routed from the source to one or several destinations without too much administrative overhead and delay.

Some of the routing protocols suitable for city environment are GSR[5], GPSR[3], A-STAR[4], VADD[8], GYTAR[2] and CAR[6]:

GPSR[3] is a position based routing which is based on two strategies: Greedy Forwarding and Perimeter Forwarding. Greedy Forwarding uses closest neighbor's information of destination in order to forward packet. This method fails when there are no closer neighbours available near the destination. Perimeter forwarding is used when greedy forwarding fails. Perimeter forwarding uses nodes in the void regions and forwards packets towards destination using right hand rule[3]. Greedy forwarding is not suitable for the vehicular networks because the nodes are highly mobile and the nodes may not be able to maintain its next hop neighbors information as the other node may go out of range due to high mobility. This can lead to data packets loss.

GSR[5] routing was proposed to deal with challenges faced by GPSR in city environment. It uses a static street map and location information about each

node. It combines position-based routing with topological knowledge by computing the sequence of intersections using Dijkstra's shortest path algorithm to reach to the destination. It uses Reactive location service for position discovery. GSR use "switch back to greedy" method for local recovery. After a packet reach to its local maximum, it switch back to greedy forwarding. This protocol neglects the situation like sparse network where there are not enough nodes for forwarding packets.

Similar to GSR, A-STAR[4] computes the number of junctions to reach the destination. In addition to it, it uses traffic information and street awareness in path finding. In street awareness, A-STAR gets the anchor information according to the street map. A-STAR uses statically and dynamically rated maps to find the number of junctions. In statistically rated maps, A- STAR uses schedule of buses to ensure the high connectivity e.g. some streets are served by regular city buses and their connectivity can be high due to presence of city buses. In dynamically rated maps, A-STAR collects the latest information of traffic to find the anchors/junctions to compute the path. When a packet face problem to pass from a junction, that junction is marked as "out of service" so other packets are restricted to traverse that junction until that junction changed to "Operational" state[4].

GyTAR[2] is an intersection-based routing scheme that finds a sequence of intersections between source and destination considering parameters such as the remaining distance to the destination and the variation in vehicular traffic. GyTAR selects the junctions dynamically. The data forwarding between the intersections adopts either an improved greedy forwarding mechanism or a carry-and-forward mechanism, depending upon the availability of the forwarding node. A score is given to each junction considering the traffic density and the curvemetric distance to the destination. The best destination junction j is then the junction with the highest score. But the neighbor table overhead is high in highly dynamic environment. The recovery strategy[2] of GyTAR is based on "carry and forward" technique.

VADD[8] is a vehicular routing that exploits the availability of map information. This routing protocol, aimed at sparsely connected vehicular networks, uses a store and forward technique and approaches the destination by selecting the direction with the lowest estimated delay to the destination. VADD has three packet modes[8]: Intersection, Straightway, and Destination based on the location of the packet carrier. By switching between these packet modes, the packet carrier takes the best packet forwarding path. Among the three modes, the Intersection mode is the most critical and complicated one, since vehicles have more choices at the intersection. The drawback of this routing protocol is that it causes large delay due to the changes of topology and traffic.

CAR[6] is a position based routing protocol which has the ability to maintain a cache of successful routes between various source and destination pairs. Initially the source broadcasts request messages to probe the destination. The request message caches the change of direction information and gathers the connectivity and hop count information. On receiving request message, the destination decides the routing path and replies to the source. Then, the data packets are forwarded along the path, as suggested by the destination. "Guards" helps to keep track of the current position of a destination, even if it travels a substantial distance from its initially known location. There are two types of guards namely standing guards and travelling guards. A standing guard represents the temporary state information that is tied to a geographical area, rather than to a specific node. The guard is kept alive by the nodes located in the area. A travelling guard also contains a velocity vector, in addition to the guarded position and radius. Each node that receives this guard records the time when the guard was received. The node computes the new guarded position based on the old guarded position, the velocity vector of the guard, and the time passed since this guard was received.

4. PROPOSED SYSTEM:

The back-bone-assisted hop greedy routing (BAHG)[7] is a position-based routing protocol used for VANET's city environment. It finds a routing path consisting of the minimum of intermediate intersections. The protocol is designed considering certain features in a city map, such as road segments, intersections, etc. A request-reply scheme is used to obtain destination position, which is then used to compute the routing path. An update keeps track of the movement of source as well as destination. Overall, the objective of the hop greedy routing algorithm is to reduce the hop count, which ultimately reduces the end-to-end delay. Figure 1 shows the backbone nodes at intersections and road segments.

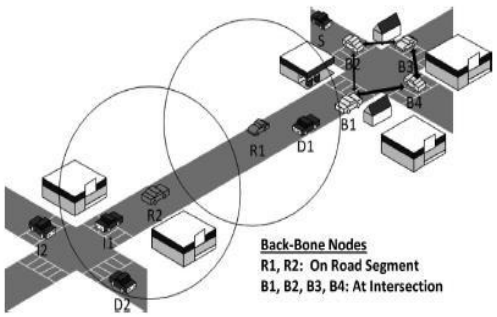


Fig 1: Back bone nodes at the intersections and road segment

Zone formation and Backbone nodes setup:

The city map is divided into several zones and some of the intersections are chosen to be the boundary intersections that are located on the outline of a zone.

Connectivity is an important requirement for a routing protocol for reliable delivery of packets. The back bone nodes are used to maintain connectivity at the intersection and road segment. The back bone nodes at the intersections maintains connectivity and the back bone nodes at the road segments detects void regions.

The back bone nodes are classified into three types namely, stable, primary and secondary back bone nodes. A stable back-bone node is selected from the group of vehicles waiting at the intersection during red traffic signal. The vehicle closest to the intersection declares itself as the stable back bone. The primary and secondary back bones are selected from the group of vehicles crossing the intersection when the signal turns green. The primary back bone is the one located at the intersection, whereas the secondary back bone is
outside the intersection. When the current primary back-bone node leaves the intersection region, it notifies the secondary back bone to become the new primary back bone.

Two-Phase Destination Discovery:

Initially, the source finds a boundary intersection as the intermediate target in each zone. Here, the boundary intersection represents a corner intersection in a zone. The hop greedy algorithm is used to discover the best possible path to the selected boundary intersection and forwards the unicast request message.

The request messages are then spread within each zone. In this approach, unicast is preferred over the broadcast as the packets are vulnerable to collision at the intersections. The back-bone nodes carries the request messages and take the responsibility to spread within their respective zones. Figure 2 shows a flowchart of destination discovery.

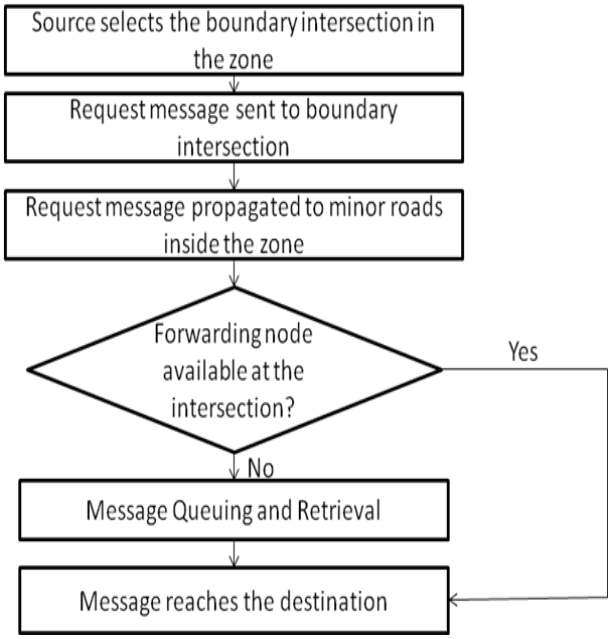


Fig 2: Two phase destination discovery

Hop greedy algorithm in Destination Reply and Data Dissemination:

Once the reply message is received from the destination, the source transmits the data packets by computing a path to the destination using the same hop greedy algorithm. The update procedure also handles the change of position of the destination if the data packets do not locate it at the old position.

Message Queuing and retrieval:

The packet buffering is performed by the stable back bone node. In the absence of forwarding node, the packet is stored in a stable back-bone node. On availability of a forwarding node in the desired direction, the packet is retrieved and forwarded. The stable back-bone nodes maintains the database of all communications with a timestamp. The stable node stores the source addresses and destination addresses along with the time of arrival of packets. If a similar packet arrives with a new timestamp, then the previous database information is updated. To allow back-bone nodes to keep track of the movements, both source and destination inform about their identity in their beacons. Whenever source or destination changes direction, the back bone node updates the corresponding entry in its communication history.

Position Update:

Before receiving the reply message, the source may change its position. The back-bone nodes

must be aware of the direction of the source movement. When a forwarder chosen among the back-bone nodes learns about such changes, it forwards the reply message toward the new direction. Ultimately, the source is able to receive the reply message. Likewise, the destination may change its position before receiving the data packet, and its movements are tracked by the back-bone nodes.

Classification of Backbone nodes:

The vehicles at the intersection are classified based on their speed to improve the reliability of routing. The slow moving vehicles like truck and bus are used as backbone nodes to improve the connectivity at the intersection.

5. CONCLUSION:

Several routing protocols are used for VANET routing in city environment. The position based routing protocols outperforms the traditional ad hoc routing protocols in VANET. In city environments, intersections play crucial roles for data communications. Thus unicast routing is preferred over the broadcast as the packets are vulnerable to collision at the intersections. As the intersection region is comparatively small and the probability of change of direction is very high, it will be risky to choose an unstable node as the forwarding node from this region. The node may cross the intersection before receiving a data packet. Thus the Back Bone assisted Hop Greedy Routing (BAHG) is used where the concept of back bone nodes avoids the connectivity issues such as void regions and unavailability of forwarders. The hop greedy algorithm finds the best possible path in terms of both hop count and connectivity. The update procedure of BAGH is also very effective to deal with the source and destination movement. The classification of nodes improve the connectivity around the intersection.

6. REFERENCES:

[1] Bijan Paul, Md. Ibrahim, Md. Abu Naser Bikas, "VANET Routing Protocols: Pros and Cons", IEEE International Journal of Computer Applications, April 2011.

[2] Jerbi.M, S. M. Senouci, T. Rasheed, and Y. Ghamri-Doudane, "Towards efficient geographic routing in urban vehicular networks," IEEE Trans. Veh. Technol., vol. 58, no. 9, pp. 5048–5059, Nov. 2009.

[3] Lee.K.C, J. Haerri, U. Lee, and M. Gerla, "Enhanced Perimeter Routing for Geographic Forwarding Protocols in Urban Vehicular Scenarios", in Proc. IEEE GlobeCom Workshops, June 2007,pp 1-10.

[4] Liu.G, B. S. Lee, B. C. Seet, C. H. Foh, K. J. Wong, and K. K. Lee,"A routing strategy for metropolis vehicular communications," in Proc. ICOIN, LNCS, Aug. 2004, pp. 134-143.

[5] Lochert . C, H. Hartenstein, J. Tian, H. Fubler, D. Hermann, and M. Mauve, "A routing strategy for vehicular ad hoc networks in city environments," in Proc. IEEE Intelligent Vehicle Symposium, June2003, pp. 156–161.

[6] Naumov.V and T. R. Gross, "Connectivity-aware routing (CAR) in vehicular ad hoc networks," in Proc. IEEE INFOCOMM, 2007, pp. 1919-1927

[7] Sahu.P.K, E. H. Wu, J. Sahoo, and M. Gerla,"BAHG: Back-Bone-Assisted Hop Greedy Routing for VANET's City Environments", IEEE transactions on intelligent transportation systems, March 2013.

[8] Zhao.J and G. Cao, "VADD: Vehicle-assisted data delivery in vehicular ad hoc networks," IEEE Transactions on Vehicular Technology., vol. 57, no. 3, May 2008, pp. 1910–1922.

[9] Schnaufer. S, "Unicast Ad-Hoc Routing in Vehicular City Scenarios", IEEE World of Wireless, Mobile and Multimedia Networks, 2008.

[10] M Nekovee, "Sensor networks on the road: the promises and challenges of vehicular ad hoc networks and vehicular grids", worksop on Ubiquitous computing and research,2005.

[11] Z. C. Taysi and A. G. Yavuz, "Routing protocols for GeoNet: A survey," IEEE Trans. Intell. Transp. Syst., vol. 13, no. 2, pp. 939–954, Jun. 2012.

[12] H. Fubler, M. Mauve, H. Hartenstein, and D. Vollmer, "A comparison of routing strategies for vehicular ad-hoc networks," Dept. Comput. Sci., Univ. Mannheim, Mannheim, Germany, Tech. Rep. TR-02-003, Jul. 2002.

[13] The Network Simulator-ns-2. [Online]. Available: http://www.isi.edu/ nsnam/ns/

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A New Algorithm for Inferring User Search Goals with Feedback Sessions

[1]K.N.Kuvalayajothi [2] N.Murugan

[1] M.Phil,(Research Scholar), 2 Assistant Professor,

[1 & 2] DEPARTMENT OF COMPUTER SCIENCE, BHARATHIDASAN COLLEGE OF ARTS AND SCIENCE,

ABSTRACT

For a broad-topic and ambiguous query, different users may have different search goals when they submit it to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. In this paper, we propose a novel approach to infer user search goals by analyzing search engine query logs. First, we propose a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. Second, we propose a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Finally, we propose a new criterion "Classified Average Precision (CAP)" to evaluate the performance of inferring user search goals. Experimental results are presented using user click-through logs from a commercial search engine to validate the effectiveness of our proposed methods.

***Keywords*—** Search engines, Web search, Search methods, Feedback, Optimization methods, Search problems, Information retrieval.

1 INTRODUCTION

In web search applications, queries are submitted to search engines to represent the information needs of users. However, sometimes queries may not exactly represent users' specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query.

For example, when the query "the sun" is submitted to a search engine, some users want to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun, as shown in Fig. 1. Therefore, it is necessary and potential to capture different user search goals in information retrieval. We define user search goals as the information on different aspects of a query that user groups want to obtain.

Information need is a user's particular desire to obtain information to satisfy his/her need. User search goals can be considered as the clusters of information needs for a query. The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience. Some advantages are summarized as follows. First, we can restructure web search results according to user search goals by grouping the search results with the same search goal; thus, users with different search goals can easily find what they want. Second, user search goals represented by some keywords can be utilized in query recommendation thus; the suggested queries can help users to form their queries more precisely. Third, the distributions of user search goals can also be useful in applications such as reranking web search results that contain different user search goals.

Due to its usefulness, many works about user search goals analysis have been investigated. They can be summarized into three classes: query classification, search result reorganization, and session boundary detection. In the first class, people attempt to infer user goals and intents by predefining some specific classes and performing query classification accordingly. End user goals as "Navigational" and "Informational" and categorize queries into these two classes. Define query intents as "Product intent" and "Job intent" and they try to classify queries according to the defined intents. Other works focus on tagging queries with some predefined concepts to improve feature representation of queries. However, since what

users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical. In the second class, people try to reorganize search results. The interesting aspects of queries by analyzing the clicked URLs directly from user click-through logs to organize search results. However, this method has limitations since the number of different clicked URLs of a query may be small. Other works analyze the search results returned by the search engine when a query is submitted. Since user feedback is not considered, many noisy search results that are not clicked by any users may be analyzed as well. Therefore, this kind of methods cannot infer user search



Fig. 1. The examples of the different user search goals and their distributions for the query "the sun" by our experiment

In this paper, we aim at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. We first propose a
novel approach to infer user search goals for a query by clustering our proposed feedback sessions. The feedback session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs.

Then, we propose a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords. Since the evaluation of clustering is also an important problem, we also propose a novel evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. We also demonstrate that the proposed evaluation criterion can help us to optimize the parameter in the clustering method when inferring user search goals. To sum up, our work has three major contributions as follows:

1) We propose a framework to infer different user search goals for a query by clustering feedback sessions. We demonstrate that clustering feedback sessions is more efficient than clustering search results or clicked URLs directly. Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered.

2) We propose a novel optimization method to combine the enriched URLs in a feedback session to form a pseudo-document, which can effectively reflect the information need of a user. Thus, we can tell what the user search goals are in detail.

3) We propose a new criterion CAP to evaluate the performance of user search goal inference based on restructuring web search results. Thus, we can determine the number of user search goals for a query.

2 FRAMEWORK OF OUR APPROACH

In the upper part, all the feedback sessions of a query are first extracted from user click-through logs and mapped to pseudo-documents. Then, user search goals are inferred by clustering these pseudo-documents and depicted with some keywords. Since we do not know the exact number of user search goals in advance, several different values are tried and the optimal value will be determined by the feedback from the bottom part.

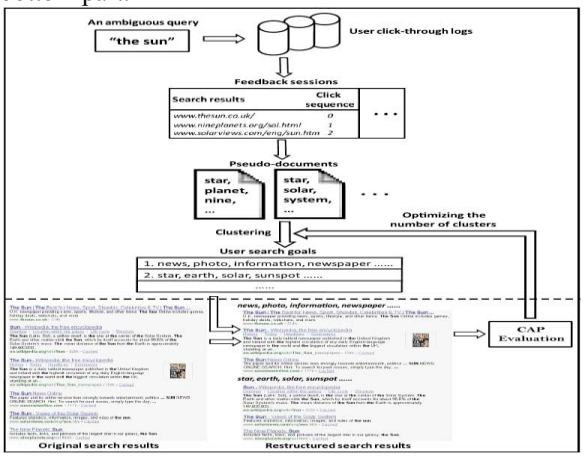


Fig. 2 shows the framework of our approach. Our framework consists of two parts divided by the dashed line.

In the bottom part, the original search results are restructured based on the user search goals inferred from the upper part. Then, we evaluate the performance of restructuring search results by our proposed evaluation criterion CAP. And the evaluation result will be used as the feedback to select the optimal number of user search goals in the upper part.

3 RELATED WORK

One application of user search is goals restructuring web search results. There are also some related works focusing on organizing the search results. In this paper, we infer user search goals from user click-through logs and restructure the search results according to the inferred user search goals.

4 CONCLUSION

In this paper, a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents. First, we introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Both the clicked URLs and the unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. Therefore, feedback sessions can reflect user information needs more efficiently. Second, we map feedback sessions to pseudo documents to approximate goal texts in user minds. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. Based on these pseudo-documents, user search goals can then be discovered and depicted with some keywords. Finally, a new criterion CAP is formulated to evaluate the performance of user search goal inference. Experimental results on user click-through logs from a commercial search engine demonstrate the effectiveness of our proposed methods. The complexity of our approach is low and our approach can be used in reality easily. For each query, the running time depends on the number of feedback sessions. However, the dimension of Ffs in (3) and (5) is not very high. Therefore, the running time is usually short. In reality, our approach can discover user search goals for some popular queries offline at first. Then, when users submit one of the queries, the search engine can return the results that are categorized into different groups according to user search goals online. Thus, users can find what they want conveniently.

REFERENCE

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.ACM Press, 1999.

[2] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf.Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[3] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416,2000.

[4] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.

[5] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li,"Context-Aware Query Suggestion by Mining Click-Through," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Resource allocation using Virtual Machines for Cloud Computing

[1]Mrs. K.Lavanya, [2]Mrs. S. Rajanandini
[1]Research Scholar, Department of Computer Science
[2]Asst. Professor. Department of Computer Application and software system
Sri Krishna Arts & Science College, Coimbatore-641008

ABSTRACT

The cloud computing focuses on maximizing the effectiveness of the shared resources. This paper aims to optimize resource use, maximize throughput, minimize response time, and avoid overload of any one of the resources. Using Virtual machines with load balancing instead of a single component may increase reliability through redundancy. Virtual machines are usually provided by dedicated software or hardware, such as a multilayer switch or a Domain Name System server process.

Keyword: Cloud computing - virtual machine - MRE – AVM - PVM - MPI

1. Introduction

Cloud computing relies on sharing of resources to achieve coherence and economies of scale, similar to a utility (like the electricity grid) over a network.[4] At the foundation of cloud computing is the broader concept of converged infrastructure and services. The cloud also focuses on maximizing the effectiveness of the shared resources. Cloud resources are usually not only shared by multiple users but are also dynamically reallocated per demand. This can work for allocating resources to users. For example, a cloud computer facility that serves European users during European business hours with a specific application (e.g., email) may reallocate the same resources to serve North American users during North America's business hours with a different application (e.g., a web server). This approach should maximize the use of computing powers thus reducing environmental damage as well since less power, air conditioning, rack space, etc. is required for a variety of functions. Proponents claim that cloud computing allows companies to avoid upfront infrastructure costs, and focus on projects that differentiate their businesses instead of infrastructure.[5] Proponents also claim that cloud computing allows enterprises to get their applications up and running faster, with improved manageability and less maintenance, and enables IT to more rapidly adjust resources to meet fluctuating and unpredictable business demand.[5]

2. Definitions

A virtual machine (VM) is a software implementation of a machine (i.e. a computer) that executes programs like a physical machine. Virtual machines are separated into two major classifications, based on their use and degree of correspondence to any real machine: A system virtual machine provides a complete system platform which supports the execution of a complete operating system (OS).[1] These usually emulate an existing architecture, and are built with the purpose of either providing a platform to run programs where the real hardware is not available for use (for example, executing on otherwise obsolete platforms), or of having multiple instances of virtual machines leading to more efficient use of computing resources, both in terms of energy consumption and cost effectiveness (known as hardware virtualization, the key to a cloud computing environment), or both.

2.1. System Virtual Machine Advantages

a. Multiple OS environments can co-exist on the same computer, in strong isolation from each other

b. The virtual machine can provide an instruction set architecture (ISA) that is somewhat different from that of the real machine

c. Application provisioning, maintenance, high availability and disaster recovery [3].

2.2. The Main Disadvantages of VMs

a. A virtual machine is less efficient than a real machine when it accesses the hardware indirectly

b. When multiple VMs are concurrently running on the same physical host, each VM may exhibit a varying and unstable performance (speed of execution, and not results), which highly depends on the workload imposed on the system by other VMs, unless proper techniques are used for temporal isolation among virtual machines.

2.3 Reason for Virtualization

In the case of server consolidation, many small physical servers are replaced by one larger physical server to increase the utilization of costly hardware resources such as CPU. Instead, each OS running on a physical server becomes converted to a distinct OS running inside a virtual machine. The large server can "host" many such "guest" virtual machines. This is known as Physical-to-Virtual (P2V) transformation.

Consolidating servers can also have the added benefit of reducing energy consumption. A typical server runs at 425 W. and VMware estimates an average server consolidation ratio of 10:1.

A virtual machine can be more easily controlled and inspected from outside than a physical one, and its configuration is more flexible. This is very useful in kernel development and for teaching operating system courses.

A new virtual machine can be provisioned as needed without the need for an up-front hardware purchase.

A virtual machine can easily be relocated from one physical machine to another as needed. For example, a salesperson going to a customer can copy a virtual machine with the demonstration software to his laptop, without the need to transport the physical computer. Likewise, an error inside a virtual machine does not harm the host system, so there is no risk of breaking down the OS on the laptop.

Examples of virtualization scenarios:

a. *Running one or more applications that are not supported by the host OS:* A virtual machine running the required guest OS could allow the desired applications to be run, without altering the host OS..

b. *Server virtualization:* Multiple virtual servers could be run on a single physical server, in order to more fully utilize the hardware resources of the physical server.

c. *Duplicating specific environments:* A virtual machine could, depending on the virtualization software used, be duplicated and installed on multiple hosts, or restored to a previously backed-up system state.

d. *Creating a protected environment:* if a guest OS running on a VM becomes damaged in a way that is difficult to repair, such as may occur when studying malware or installing badly behaved software, the VM may simply be discarded without harm to the host system, and a clean copy used next time.

3. Process Virtual Machines

A process VM, sometimes called an *Managed Runtime Environment* (MRE), or *application virtual machine(AVM),* runs as a normal application inside a host OS and supports a single process. It is created when that process is started and destroyed when it exits. Its purpose is to provide a platform-independent programming environment that abstracts away details of the underlying hardware or operating system, and allows a program to execute in the same way on any platform.

A process VM provides a high-level abstraction — that of a high-level programming language (compared to the low-level ISA abstraction of the system VM). Process VMs are implemented using an interpreter; performance comparable to

compiled programming languages is achieved by the use of just-in-time compilation.

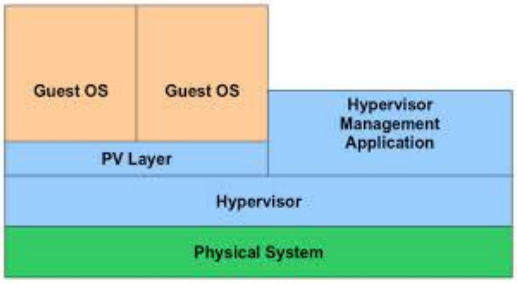
This type of VM has become popular with the Java programming language, which is implemented using the Java virtual machine. Other examples include the Parrot virtual machine, which serves as an abstraction layer for several interpreted languages, and the .NET Framework, which runs on a VM called the Common Language Runtime.

A special case of process VMs are systems that abstract over the communication mechanisms of a (potentially heterogeneous) computer cluster. Such a VM does not consist of a single process, but one process per physical machine in the cluster. They are designed to ease the task of programming concurrent applications by letting the programmer focus on algorithms rather than the communication mechanisms provided by the interconnect and the OS. They do not hide the fact that communication takes place, and as such do not attempt to present the cluster as a single machine.

Unlike other process VMs, these systems do not provide a specific programming language, but are embedded in an existing language; typically such a system provides bindings for several languages (e.g., C and FORTRAN). Examples are PVM (Parallel Virtual Machine) and MPI (Message Passing Interface). They are not strictly virtual machines, as the applications running on top still have access to all OS services, and are therefore not confined to the system model.

4. Operating System-Level Virtualization

Operating system-level virtualization is commonly used in virtual hosting environments, where it is useful for securely allocating finite hardware resources amongst a large number of mutually-distrusting users.



OS-level virtualization implementations that are capable of live migration can be used for dynamic load balancing of containers between nodes in a cluster.

Operating system-level virtualization is a server virtualization method where the kernel of an operating system allows for multiple isolated user-space instances, instead of just one. Such instances (often called containers, VE (virtual environments), VPS (virtual private servers) or jails) may look and feel like a real server, from the point of view of its owner.

Operating system-level virtualization is not as flexible as other virtualization approaches since it cannot host a guest operating system different from the host one, or a different guest kernel. For example, with Linux, different distributions are fine, but other OS such as Windows cannot be hosted.

This limitation is partially overcome in Solaris by its branded zones feature, which provides the ability to run an environment within a container that emulates an older Solaris 8 or 9 versions in a Solaris 10 host. (a Linux branded zone was also announced and implemented for some Linux kernels, but later abandoned).

4.1 Server Virtualization

Server virtualization, a server computer hosts one or more, other server computers. Server virtualization makes cloud computing feasible.

Virtual servers seek to encapsulate the server software away from the hardware. This includes the OS, the applications, and the storage for that server. Servers end up as mere files stored on a physical box, or in enterprise storage. A virtual server can be serviced by one or more hosts, and one host may house more than one virtual server.

Virtual servers can be scaled out easily. If the administrators find that the resources supporting a virtual server are being taxed too much, they can adjust the amount of resources allocated to that virtual server

Server templates can be created in a virtual environment to be used to create multiple, identical

virtual servers. Virtual servers themselves can be migrated from host to host almost at will.

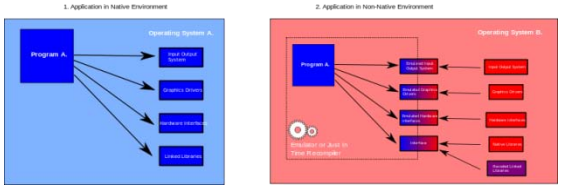
VMware vSphere gives a complete server virtualization platform that delivers:

- Server resources used grater then 80 percent

- Savings in operating costs and capital is Up to 50 percent

- 10:1 of server consolidation ratio

4.2 Application Virtualization



Application virtualization is software technology it encapsulates application software from the underlying OS on which it is executed.[1] Although it is still executed as if it were. The application behaves at runtime like it is directly interfacing with the original operating system and all the resources managed by it, but can be isolated or sandboxed to varying degrees. In this context, the term "virtualization" refers to the artifact being encapsulated (application), which is quite different from its meaning in hardware virtualization, where it refers to the artifact being abstracted (physical hardware).

Full application virtualization requires a virtualization layer. The layer intercepts all file and Registry operations of virtualized applications and transparently redirects them to a virtualized location; often a single file.

Limitations of application virtualization:

a. All software cannot be virtualized. Some examples include applications that require a device driver and 16-bit applications that need to run in shared memory space.[8]

b. Some types of software such as anti-virus packages and applications that require heavy OS integration, such as Stardock's WindowBlinds or TGT Soft's StyleXP are difficult to virtualized.

c. Only file and registry-level compatibility issues between legacy applications and newer operating systems can be addressed by application virtualization. For example, applications that don't manage the heap correctly will not execute on Windows Vista as they still allocate memory in the same way, regardless of whether they are virtualized or not.[1]

d. Moreover, in software licensing, application virtualization bears great licensing pitfalls mainly because both the application virtualization software and the virtualized applications must be correctly licensed.[3]

4.3 Storage Virtualization

Storage virtualization is part of the software-defined storage layer that must offer improvements in performance and space efficiency without requiring the purchase of additional storage hardware.It must enable rapid provisioning so that high-performance, space-efficient storage can be spun up as fast as a VM can be spun up today. It must offer a VM-centric storage management model that is intuitive for virtual administrators who are taking on more of the storage management tasks in virtual environments.VMware storage virtualization is a combination of capabilities that provide an abstraction layer for physical storage resources to be addressed, managed and optimized in a virtualization deployment.

Storage virtualization technology provides a fundamentally better way to manage storage resources for your virtual infrastructure, giving your organization the ability to:

a. Significantly improve storage resource utilization and flexibility

b. Simplify OS patching and driver requirements, regardless of storage topology

c. Increase application uptime and simplify day-to-day operations

d. Leverage and complement your existing storage infrastructure

5. VMware Cloud Solutions

VMware cloud solutions improve IT efficiency; agility and reliability, while helping IT drive innovation. We deliver everything IT needs to build, operate, rent and manage their cloud, even beyond the data center, while continuously quantifying its impact. VMware cloud solutions maximize the potential of cloud computing to:

a. *Deliver new IT services that fuel business growth* – More rapidly create and deploy services that differentiate the business.

b. *Transform IT into a source of innovation* – Unlock IT resources and reinvest in services that advance business goals.

c. *Ensure IT efficiency, agility and reliability* – Deliver enterprise SLAs for tier 1 applications, securing the business even across diverse cloud environments.

Running virtual machines on Windows Azure combines the benefits of virtualization and cloud computing.

Cloud computing has taken that degree of efficiency and agility realized from virtualization and magnified it further. Through pooled resources, geographic diversity and universal connectivity, cloud computing has facilitated delivering hosted software, platforms and infrastructure as a service. With the introduction of the virtual machine (VM) role on the Windows Azure platform, these powerful concepts come together to host virtual infrastructure in the cloud.

Building for the Cloud

The VM role for Windows Azure is similar to other server roles. VMs on Windows Azure run a version of the Windows Server OS.

Install the Windows Azure Integration Components. These are required on every server image before you can upload it to Windows Azure.

Developing an "adapter" is an optional next step in the image-build process. When you develop and upload a custom server image, you're installing and configuring software to run in the dynamic environment of Windows Azure.

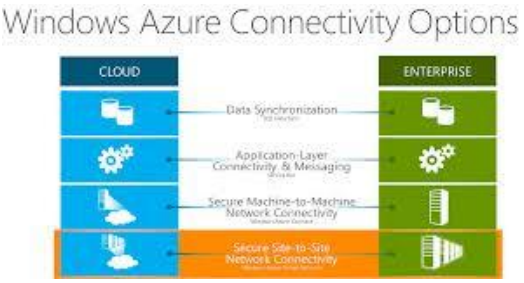
You have two options for writing an adapter:

a. You can write an adapter that runs during the specialization pass. This occurs while the OS is being set up, after the server image is uploaded to Windows Azure for the first time or after an instance is reimaged.

b. You can write an adapter as a Windows service—one that starts automatically each time the OS starts. You can write this type of adapter in either managed or native code.

One final step to prepare your system image to be uploaded to Windows Azure is to evaluate your Windows Firewall configuration. You'll want to open any ports your application needs after it's running on the VM role instance in Windows Azure.

5.1 Deploying a VM to Windows Azure



The first step in deploying your system image is to configure management and service certificates. When you upload a .vhd file to Windows Azure, it uses an x.509 certificate, saved as a .cer file, for authentication. You also need a Personal Information Exchange certificate, saved as a .pfx file (referred to as a service certificate in the Windows Azure Management Portal), if you plan to remotely access the VM role instance.

6. Conclusion

In this paper we discus about the importance of resource allocation in cloud computing. Virtual

machines will copy server details and share and it will be much cost efficient, it doesn't need more memory space. VMware cloud solutions improve IT efficiency. Multiple virtual servers could be run on a single physical server, in order to more fully utilize the hardware resources of the physical server.

7. References

[1] K. Chard and K. Bubendorfer, "Using Secure Auctions to Build A Distributed Meta-cheduler for the Grid," Market Oriented Grid and Utility Computing, series Wiley Series on Parallel and Distributed Computing, R. Buyya and K. Bubendorfer, eds., pp. 569-588, Wiley, 2009.

[2] V. Vinothina, Dr. R. Shridaran, and Dr. Padmavathi Ganpathi, A survey on resource allocation strategies in cloud computing, International Journal of Advanced Computer Science and Applications, 3(6):97-104,2012.

[3] Gunho Lee, Niraj Tolia, Parthasarathy Ranganathan, and Randy H.Katz, Topology aware resource allocation for data-intensive workloads, ACM SIGCOMM Computer Communication Review, 41(1):120--124, 2011.

[4] Daniel Warneke and Odej Kao, Exploiting dynamic resource allocation for efficient parallel data processing in the cloud, IEEE Transactions On Parallel And Distributed Systems, 2011.

[5] Cloud Computing architecture http://blog.infizeal.com/2011/12/cloud -computing architecture .html

[6] Cheng Shiwei, Pan Yu.Credibility-based dynamic resource distribution strategy under cloud computing environment [J]. Computer Engineering, 2011, 6(37):45~48.

[7] Siva Theja Maguluri and R. Srikant
Department of ECE and CSL" Heavy Traffic Optimal Resource Allocation Algorithms for Cloud Computing Clusters"

[8] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing sla violations," in Proc. of the IFIP/IEEE International Symposium on Integrated Network Management (IM'07),
2007.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

AN EFFICIENT METHOD FOR THE ESTIMATION AND EXTRACTION OF RGB COLOR DATA IN A MASKED IMAGE USING SCATTER PLOT

D. Napoleon[1], V.Mageshwari[2], K.Prabavathi [3]

[1] Assistant Professor, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, India, Coimbatore-641046

[2 3] Research Scholars, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, India, Coimbatore-641046

ABSTRACT

Images are clones of real world substances. Images are referenced to the digital data in the real world. Color Model plays a predominant role in Image Processing. RGB color model is an additive color model because red, green, and blue light are added together in various ways to reproduce a broad array of colors in RGB color model. In this research paper RGB color image model has been used which has been wrapped using conformal transformation. One of the important properties of conformal transformation is the preservation of local shape in Images. Masking refers to the transferring of particular areas into another background which refers to changing the color of the specific areas of a picture. In this research work first the objects of transformed RGB color model image is wrapped with binary 1's and 0's bit mask. Then the transformed image is wrapped with a 0's bit mask. The color data in the image masked with binary 1's and 0's bits are completely hidden but the color data in the image masked with 0's bit is not completely hidden. It gives some information about the color of the input image. Finally scatter plots are plotted for input and the output images. It shows that the color data is not totally hidden but it is only limited in the 0's bit masking.

Keywords: Conformal Transformation; Image Masking

1. INTRODUCTION

RGB colors are called as additive colors because they are formed using combination of red, green and blue colors in various proportions. RGB is a device-dependent color model. The RGB color model is used for sensing, representing, and displaying of images in electronic systems, such as computers and television. Conformal transformation has many properties and uses. Preservation of local shapes is one of the appropriate properties among them. It is also used to preserve angles. Masking is one of the important concepts in image processing. Image masking is used to mask the picture. During the process of masking, parts of the image will be altered so much that they will become unrecognizable. In digital image processing, masking refers to the transferring of particular areas into another background or it also refers to changing the color of the specific areas of a picture. It can also be used to hide the particular portion of a picture. Mainly masking concept is to hide the image data partially or totally.

Binary masks are created by changing specific bits in the original value to the desired setting. Using a Boolean comparison operation it is also used to

change one or more bits from 1 to 0 or vice versa. Masking of RGB image with 0's bit gives at least an idea about the color of the original image. But the data in the masked image will be less than the original image i.e., only limited data will be in the masked image. A scatter plot, scatter plot or scatter graph is a type of mathematical representation using Cartesian coordinates to display values of two variables for a set of data. The 3D scatter plots are used to plot data points on three axes in the endeavor to shows the association between three variables. The relation between dissimilar variables is called correlation.

2. METHODOLOGY:

The system architecture of our proposed work is shown in the fig.1

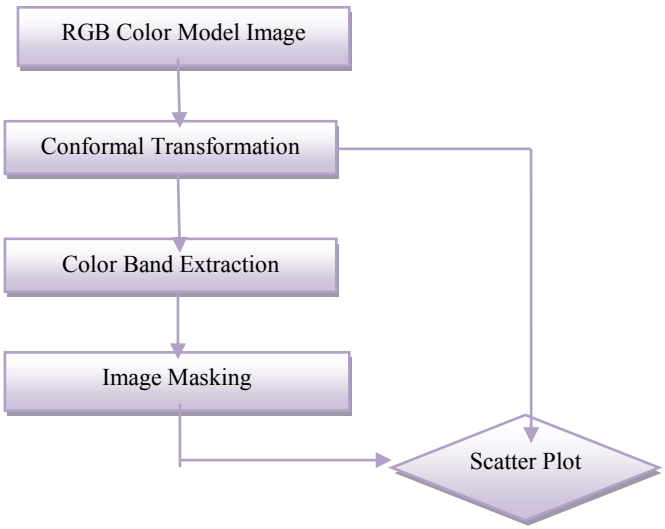


Fig1. System Architecture

3. RGB IMAGES

Red, green, and blue light are added together in various ways to reproduce a broad array of colors in RGB color model. The initials of the three additive primary colors red, green, and blue forms the name of the model[1]. The RGB color model is used for sensing, representing, and displaying of images in electronic systems, such as computers and television. It is also been used in conventional photography. The RGB color model already had a theory behind it based in human perception of colors before the electronic age. RGB is a device-dependent color model. The given RGB value is detected differently in different devices. Thus its value does not describe the same color across devices without some kind of color management.

4. CONFORMAL TRANSFORMATION

There are many important properties and uses of conformal transformation. One property relevant to image transformation is the preservation of local shape. It is also used to preserve angles.In this paper 2-D conformal transformation is used to warp an image[10]. The mapping from output to input, g: R^2 -> R^2, is defined in terms of a complex analytic function G: C -> C, where

$$G(z) = (z + 1/z) / 2.$$

We define g via a direct correspondence between each point (x,y) in R^2 (the Euclidean plane) and the point z = x + i*y in C (the complex plane),

$$g(x,y) = (Re(w),Im(w)) = (u,v)$$

where,

$$w = u + i*v = G(x + i*y)$$

We apply imtransform using the SIZE parameter to ensure an aspect ratio that matches the proportions in xData and yData (6/5), and view the result. The original and transformed images are compared. In the transformed image the edges are curved but the image is preserved by the transformation. Each feature from the original image is shown twice in the transformed image. There is also a hole found in the center of the transformed image with four regular cups around its edges to preserve images[12].

5. MASKING

Masking concept is very important in image processing. In digital image processing, masking refers to the transferring of particular areas into another background or it also refers to changing the color of the specific areas of a picture. The first step of masking process is clipping the relevant areas. The editor uses various tools to clip a specific shape or color gradient. Even manual selection with the aid of a graphic tablet or a cursor is possible with most image editing programs. A fulfilled selection then enables the masking

process. During the process of masking, parts of the image will be altered so much that they will become unrecognizable[4].

In a related way, the editor can also mask images when certain objects in the photography should not be visible. In this process the area of the object is masked first. This method works easily for masking. In this paper, a transformed RGB color model image is taken. It is first individually masked with binary 1's and 0's bits. In the next step the total objects are binary masked. In final step the image is masked using 0's bit alone.

5.1 Binary Mask

Binary masks are created by changing specific bits in the original value to the desired setting. Using a Boolean comparison operation it is also used to change one or more bits from 1 to 0 or vice versa. In computer graphics, for each proposed image there are two bitmaps: an original image in which the real areas are given a pixel value of all bits set to 0s and the neighboring areas a value of all bits set to 1s. All-zero bits contribute black pixels and all-one bits contribute white pixels. The binary mask creates an entire mapping on the image so that the original image is fully covered by black and white mask and the color of the original image is unrecognizable[3]. The binary masking gives no idea about the color of original image. Masking of RGB image with the 0's bit gives at least an idea about the color of the original image. But the data in the masked image will be less than the original image i.e., only limited data will be in the masked image.

6. SCATTER PLOT

A scatter plot or scatter graph is a type of mathematical representation using Cartesian coordinates to display values of two variables for a set of data. The data is displayed as a set of points, each having the value of one variable shaping the location on the horizontal axis and the value of the other variable shaping the location on the vertical axis. The kind of plot is called as the scatter gram, scatter diagram or scatter graph. A three-dimensional scatter plots displays the connection between three continuous variables at once. The data are represented as a cloud of points within the three-dimensional space formed by the three axes of the plot[6]. Such plot can be complex to interpret and other methods of representing the third variable are generally more efficient. The scatter plot is formed by the following function

plot3(rB(:),gB(:),bB(:),'.')

The 3D scatter plots are used to plot data points on three axes in the endeavor to shows the association between three variables. The relation between dissimilar variables is called correlation.

One of the greatest aspects of a scatter plot, however, is its skill to show nonlinear relationship between variables.

masking is used to hide the data completely. The final picture is masked only by the

0's bits and it at least gives some identification about the color of the image. Scatter plots exactly shows the details of the color data. The binary masking with 1's and 0's bits completely hides the color data of image but the masking done only with 0's bits gives some information about the color data.

7. CONCLUSION

An RGB color image model is transformed using conformal transformation. The transformed image is binary masked with 1's and 0's bits in such a way that the color of the image is invisible and it makes the image unrecognizable. This kind of

REFERENCE

[1] A.C. Hardy, F.L. Wurzburg, The theory of tri-color reproduction, J.Opt.Soc.Am. 27,1937, 227-240.

[2] R. Schinzinger, P. A. A. Laura. Comformal Mapping: Methods and Applications. Oxford: Elsevier,1991.

[3] N. Li and P.C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Immplications for noise reduction," JASA, vol.123, no. 3,pp. 1673-1682, 2008.

[4] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis,' in speech Separation by Humans and Machines, Pierre Divenyi, Ed., pp. 181-197. Kluwer, 2005.

[5] C. Frederick and E. L. Schwartz. Conformal image warping. IEEE Computer Graphics and Applications, March, 1990:54 61, 1990.

[6] P. S. Heckbert. Survey of texture mapping. IEEE Computer Graphics and Applications, 6(11):56 67, November 1986.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

[8] G. RAMPONI, N. STROBEL, S. K.MITRA, AND T. YU, "nonlinear unsharp masking methods for image contrast enhancement," J. Electron. Image, vol. 5, pp. 353–366, July 1996.

[9] H. R. Sheikh and A. C. Bovik. Image information and visual quality. IEEE Transactions on Image Processing, vol. 15, no.2, pp. 430-444, 2006.

[10] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment, IEEE Transactions on Image Processing, vol. 20, no. 5, pp. 1185-1198, 2011.

[11]H.R. Sheikh, M.F. Sabir, and A.C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms", IEEE Trans. IP, vol. 15, pp. 3440-3451, 2006.

[12]Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Trans. IP, vol. 13, pp. 600-612, 2004.

[13] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multi-scale structural similarity for image quality assessment", ACSSC'03,

2003.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Traitor tracing and secure transmission of files in networks

[1]S.Vasanth Kumar [2] N.Murugan
[1] M. Phil (Computer Science) Research Scholar, [2]AssistantProfessor
Bharathidasan College of Arts and Science, Erode.

ABSTRACT

Traitor-tracing technique using in the rapid advance in broadband technology, streaming technology is applied to many applications, such as content delivery systems and web conference systems. On the other hand, we must implement digital rights management (DRM) to control content spreading and to avoid unintended content use. Traitor tracing is one of the key technologies that constructs DRM systems, and enables content distributors to observe and control content reception. General methods make use of watermarking to provide users' individual information unique to each user. However, these methods need to produce many individual contents. Especially, this is not realistic for real-time streaming systems. Furthermore, watermarking, which is a key technology adopted by contemporary methods, has known limitations and attacks against it. This is why the authors have proposed a method to monitor the content stream using traffic patterns constructed from only traffic volume information obtained from routers. The proposed method can determine who is watching the streaming content and whether or not a secondary content delivery exists. This information can be also used with general methods to construct a more practical traitor-tracing system. A method to cope with random errors and burst errors has also been investigated. Finally, the results of simulation and practical experiment are provided demonstrating the effectiveness of the proposed approach.

Index Terms—Digital rights management (DRM), pattern recognition, streaming contents, traffic pattern, traitor tracing.

I. INTRODUCTION

In recent years, with the rapid advance in broadband technology, streaming technology has been applied for many applications, such as content delivery systems, video conference systems, e-learning systems, remote diagnosis systems, and so on Video conference systems, which use local area networks (LANs) and the Internet, are a new technology and replace previous systems which use leased lines. For example-learning is already used for educational purposes in companies. The most important feature of this kind of system that a person can obtain many kinds of lectures at any time without inviting instructors. This increases the efficiency of selflearningand reduces the cost of the education. However, these useful streaming systems can be threatened by many security problems. An example of these problems is the use of content without knowledge of the content holders and content providers. Such uncomprehendeduses involve retransmission of content as well as unauthorized uses. Hence, there is a high expectation on the digital rights management (DRM) technology which manages content usage.

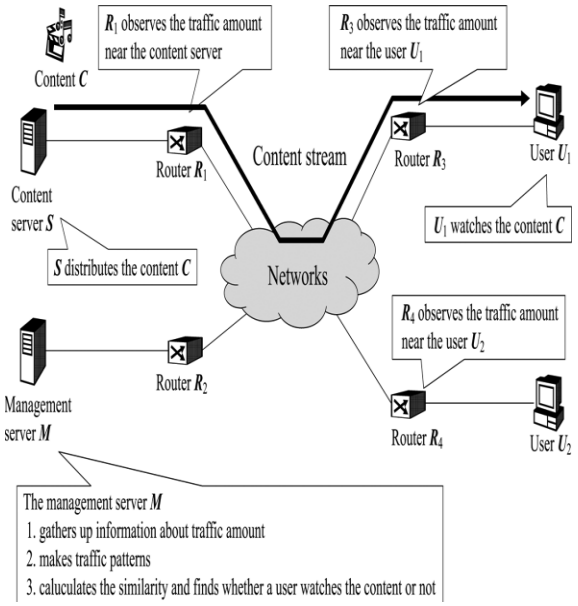
II.EXISTING SYSTEM

General traitor-tracing methods use digital watermarks and encryption keys to observe propagation of the content. The concept of general traitor tracing is to assign content with an individual identity. Ideally, watermarks embedded in the content are different from one other. However, this method requires a large amount of computations to encode many contents and to embed many watermarks. Since these costs are critical for real-time streaming, the reduction of these computations has been attempted in and addition to the computation, the shortcomings of watermarking, such as resistance to bit errors and the conversion of contents, are also serious issues. Furthermore, there are also known attacks against watermarks and cracking methods for decoders. An example of these attacks is the removal attack, which ruins watermarks by adding noise into the contents. Generally, an idealized DRM technology can continually manage content delivery and user's operations throughout the content's commercial life cycle. In the protection of content with DRM technology, content is encrypted, and decryption keys are securely transmitted to users. Users can extract the content from capsules using received keys, which avoid data interception. However, since this kind of protection is not able to restrain distributions of decrypted contents or use of plagiarized keys, content providers also implement the active management technology to monitor the content usage. This

technology is called traitor-tracing technology, which enables content providers to track the content usage.

III.PROPOSED WORK

The Bit rate of variable bit rate (VBR) streaming content changes corresponding to changes in the content. When the content is delivered to the user as a streaming content, the traffic amount observed at routers near the server and the user represents a pattern unique for each content just as human fingerprints. By comparing user-side patterns to server-side ones, we are able to determine whether or not a user is watching the targeted content. The following is a brief procedure to determine whether or not a user is watching the content. First, a server-side pattern of the streaming content is generated based on the observation of the traffic amount at the router near the content server. Next, a user-side pattern is generated in the same manner. Finally, the management server compares the user-side pattern with the server-side one to make judgments on the content reception of the user. This process requires no operation on the user's computer. In principle, this makes it impossible for the user to tamper with the tracing process. Additionally, since the proposed method makes use of only the information about traffic amount, it can be implemented with a lower computational cost than traditional methods which use watermark technology.



IV.NETWORK BASED STEAMING CONDENT USING TRAITOR TRACING

A. Definition of Traffic patterns

We focus on VBR streaming content and make use of a "traffic pattern" or the fluctuation of traffic amount, to find the streaming content flow. The traffic pattern is defined as the amount of traffic for certain
Time-slot and expressed as dimension vector Here, is the whole length of the traffic pattern and the number of "time slots. In the proposed method, the server-side router observes the traffic amount during the entire time that the content is being delivered and the server-side traffic pattern.

B. Comparison of Traffic Patterns

The outline of comparison of traffic patterns is shown in the traffic patterns consists of three steps are we place a window , which snips off a partial pattern from the server-side traffic pattern These three steps are conductedbeforecalculatingsimilarity. When there is no burst error, these processes do not cause any change on traffic patterns.

C. Dynamic Detection in Threshold

Packet loss and jitter influence the proposed method. For instance, if packets drop in the middle of content streaming, the user-side traffic pattern will deteriorate totally or partly, so that the maximum value of similarity becomes smaller. Likewise, jitter deteriorates the user-side traffic pattern. Thus, in this case, it is difficult to find a peak in the similarity graph with the static determination threshold. That is why, to accommodate the proposed method for a changing network environment, a dynamic
Threshold is required.

V.MODULE DESCRIPTION

A. Server side streaming monitoring:

In the server module, server-side pattern of the streaming content is generated based on the observation of the traffic amount at the router near the content server. This content server contains all the process about the clients. Also this is used to monitor all the client's contents from server side. In this process, each content should be involved in the pattern matching process.

B. Management server:

The management server compares the user-side pattern with the server-side one to make judgments on the content reception of the user. This process requires no operation on the user's computer. In principle, this makes it impossible for the user to tamper with the tracing process. To improve the success rate of tracking traitors, we need to discuss a

technique to narrow the list of users who may be traitors by monitoring content distribution. As well as using alone, this technique aims to assist the general traitor tracing method.

***C.* Client side streaming process:**

In the client side pattern also same as the server side streaming process .This also includes the pattern matching which identifies the original from the fake one. Then the original data send to the user. Then the user login into the client system and receive the original message successfully.

VI.CONCLUSION

We have to figure out how to protect the confidential information from being accessed by attackers. From a commonsense perspective, we must deploy security features such as encryption, access control, and so forth. Of course, using watermarking schemes for streaming contents is one of the most popular ideas of implementing security. In addition, packet monitoring for the communication that takes place between routers may also be a practical solution. Our proposed method falls into this second category. If there are attackers who have exploited knowledge of our system, they may resort to changing the bit-rate of the content. The experimental results demonstrate that our scheme does not depend on bit-rates. Furthermore, the most effective way an attacker can manipulate against our system is that the VBR content is changed into a constant bit-rate stream. This is analogous to copying the content into a USB memory and bringing it out from the server illegally. Indeed, this is a focus of physical intrusion detection mechanisms and we are, unfortunately, unable to solve this problem by using network technologies.

Future enhancements:

The future work is to advance the determination accuracy. The segmentation method of packets and the measurement of similarity of traffic patterns may be further improved.

VII. REFERENCES

[1] Z. Yang, H. Ma, and J. Zhang, "A dynamic scalable service model forsip-based video conference," in *Proc. Ninth Int. Conf. Computer Supported Cooperative Work in Design*, May 2005, vol. 1, pp. 24–26.

[2] M. Shimakawa, D. P. Holed, and F. A. Tobago, "Video-conferencing and data traffic over an ieee 802.11g wlan using dcf and edca," in*Proc.Int.Conf.Communications (ICC)*, May 2005, vol. 2, pp. 16–20.

[3]. O. Berkman, M. Parnas, and J. Sgall. E±cient dynamic traitor tracing. to appear at SODA 2000, 2000.

[4].Y. Nishimoto, A. Baba, T. Kurioka, and S. Namba, "A digital rights management system for digital broadcasting based on home servers,"*IEEE Trans. Broadcast.*, vol. 52, no. 2, pp. 167–172, Jun. 2006.

[5].B. Chor, A. Fiat, M. Naor, and B. Pinkas, "Tracing traitors," *IEEE*
Trans. Inf. Theory, vol. 46, no. 3, pp. 893–910, May 2000.

[6].S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung, "Resolving rightfulownerships with invisible watermarking techniques: Limitations, attacks,and implications," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 4,pp. 573–586, May 1998.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Data mining and Clustering

[1]Pradeep.K [2]Meena Preethi.B

[2]Assistant professor (CA & SS)

Sri Krishna Arts and Science College, Coimbatore.

ABSTRACT

The data-mining step in knowledge discovery is computationally expensive when working with large data-sets. Clustering is one form of unsupervised learning used in data mining.

INTRODUCTION

WHAT IS DATA MINING???

Data mining in general is the search for hidden patterns that may exist in large databases.the practice of examining large pre-existing databases in order to generate new information.Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

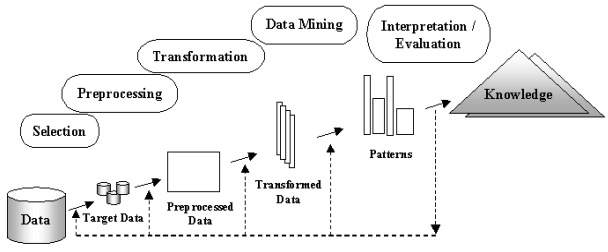


Figure 1. The Data Mining Process (from [4] pg 10)

Continuous Innovation

Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

EXAMPLE

For example, one Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure beer and diapers were sold at full price on Thursdays.

DATA MINING CONSISTS OF FIVE MAJOR ELEMENTS:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

DIFFERENT LEVELS OF ANALYSIS ARE AVAILABLE:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

WHAT TECHNOLOGICAL INFRASTRUCTURE IS REQUIRED?

Today, data mining applications are available on all size systems for mainframe, client/server, and PC platforms. System prices range from several thousand dollars for the smallest applications up to $1 million a terabyte for the largest. Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes. NCR has the capacity to deliver applications exceeding 100 terabytes. There are two critical technological drivers:

- **Size of the database**: the more data being processed and maintained, the more powerful the system required.

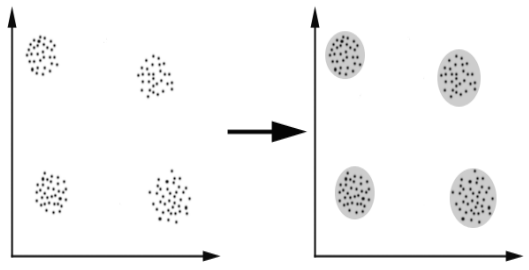
- **Query complexity**: the more complex the queries and the greater the number of queries being processed, the more powerful the system required.

Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes. However, this infrastructure needs to be significantly enhanced to support larger applications. Some vendors have added extensive indexing capabilities to improve query performance. Others use new hardware architectures such as Massively Parallel Processors (MPP) to achieve order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers.

CLUSTERING: AN INTRODUCTION

WHAT IS CLUSTERING?

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. We can show this with a simple graphical example:



In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance).

This is called *distance-based clustering*. Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

THE GOALS OF CLUSTERING

So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (*data reduction*), in finding "natural clusters" and describe their unknown properties (*"natural" data types*), in finding useful and suitable groupings (*"useful" data classes*) or in finding unusual data objects (*outlier detection*).

POSSIBLE APPLICATIONS

Clustering algorithms can be applied in many fields, for instance:

- *Marketing*: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- *Biology*: classification of plants and animals given their features;
- *Libraries*: book ordering;
- *Insurance*: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- *City-planning*: identifying groups of houses according to their house type, value and geographical location;
- *Earthquake studies*: clustering observed earthquake epicenters to identify dangerous zones;
- *WWW*: document classification; clustering weblog data to discover groups of similar access patterns.

REQUIREMENTS

The main requirements that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- high dimensionality;
- interpretability and usability.

PROBLEMS

There are a number of problems with clustering. Among them:

- current clustering techniques do not address all the requirements adequately (and concurrently);
- dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
- the effectiveness of the method depends on the definition of "distance" (for distance-based clustering);
- if an *obvious* distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces;

the result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

THE CLUSTERING IN DATA MINING….

Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Type of data in clustering analysis

- Interval-scaled variables

- Binary variables
- Nominal, ordinal, and ratio variables
- Variables of mixed types

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in four steps:
  - Partition objects into *k* nonempty subsets
  - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when no more new assignment
- Given *k*, the *k-means* algorithm is implemented in four steps:
  - Partition objects into *k* nonempty subsets
  - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when no more new assignment
- PAM (Kaufman and Rousseeuw, 1987), built in Splus
- Use real object to represent the cluster
  - Select ***k*** representative objects arbitrarily
  - For each pair of non-selected object ***h*** and selected object ***i***, calculate the total swapping cost TC_{ih}
  - For each pair of ***i*** and ***h***,
    - If $TC_{ih} < 0$, ***i*** is replaced by ***h***
    - Then assign each non-selected object to the most similar representative object
  - repeat steps 2-3 until there is no change

WHAT IS THE PROBLEM WITH PAM?

- Pam is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean
- Pam works efficiently for small data sets but does not **scale well** for large data sets.
  - $O(k(n-k)^2)$ for each iteration

  where n is # of data,k is # of clusters

➔ Sampling based method,

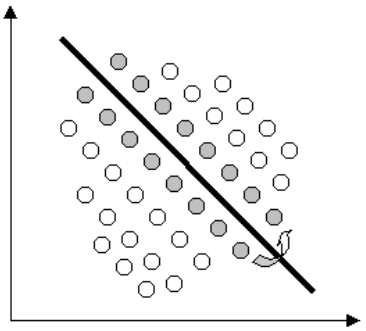
CLARA(Clustering LARge Applications)

Clustering Algorithms

Classification
Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. A simple example of that is shown in the figure below, where the separation of points is achieved by a straight line on a bi-dimensional plane.
On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.



Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Finally, the last kind of clustering use a completely probabilistic approach.

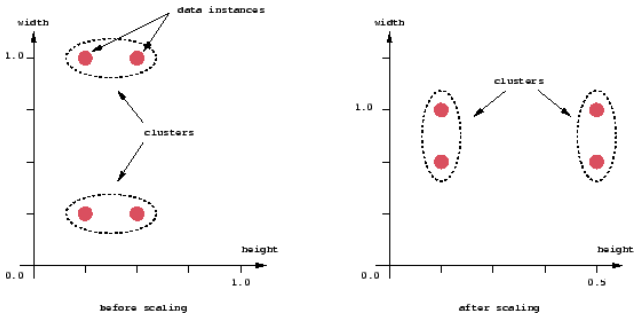
four of the most used clustering algorithms:

- K-means
- Fuzzy C-means
- Hierarchical clustering
- Mixture of Gaussians

Each of these algorithms belongs to one of the clustering types listed above. So that, K-means is an *exclusive clustering* algorithm, Fuzzy C-means is an*overlapping clustering* algorithm, Hierarchical clustering is obvious and lastly Mixture of Gaussian is a *probabilistic clustering* algorithm. We will discuss about each clustering method in the following paragraphs.

Distance Measure

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances. However, even in this case the Euclidean distance can sometimes be misleading. Figure shown below illustrates this with an example of the width and height measurements of an object. Despite both measurements being taken in the same physical units, an informed decision has to be made as to the relative scaling. As the figure shows, different scalings can lead to different clusterings.



Notice however that this is not only a graphic issue: the problem arises from the mathematical formula used to combine the distances between the single components of the data feature vectors into a unique distance measure that can be used for clustering purposes: different formulas leads to different clusterings.
Again, domain knowledge must be used to guide the formulation of a suitable distance measure for each particular application.

Minkowski Metric

For higher dimensional data, a popular measure is the Minkowski metric,

$$d_p(x_i, x_j) = \left(\sum_{K=1}^{d} \left| x_{i,k} - x_{j,k} \right|^p \right)^{\frac{1}{p}}$$

where *d* is the dimensionality of the data. The *Euclidean* distance is a special case where *p*=2, while *Manhattan* metric has *p*=1. However, there are no general theoretical guidelines for selecting a measure for any given application.

It is often the case that the components of the data feature vectors are not immediately comparable. It can be that the components are not continuous variables, like length, but nominal categories, such as the days of the week. In these cases again, domain knowledge must be used to formulate an appropriate measure.

FUZZY C-MEANS CLUSTERING

The Algorithm

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad , \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}} \quad ,$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

The algorithm is composed of the following steps:

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
2. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

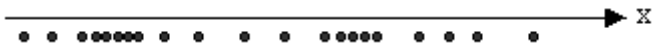
3. *Update* $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}}$$

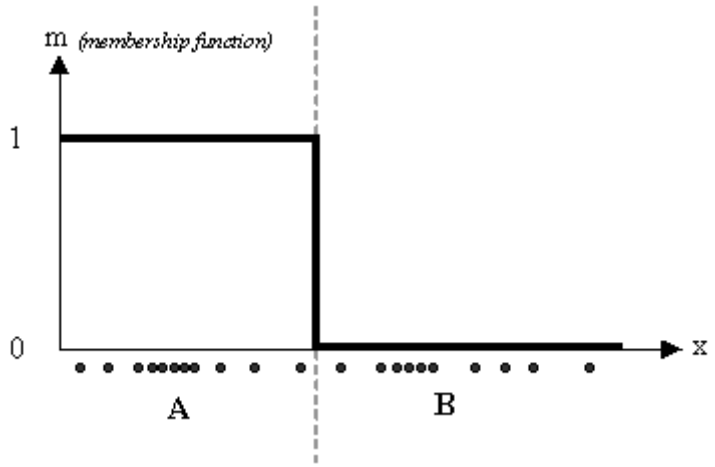
4. *If* $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ *then STOP; otherwise return to step 2.*

CONCLUSION

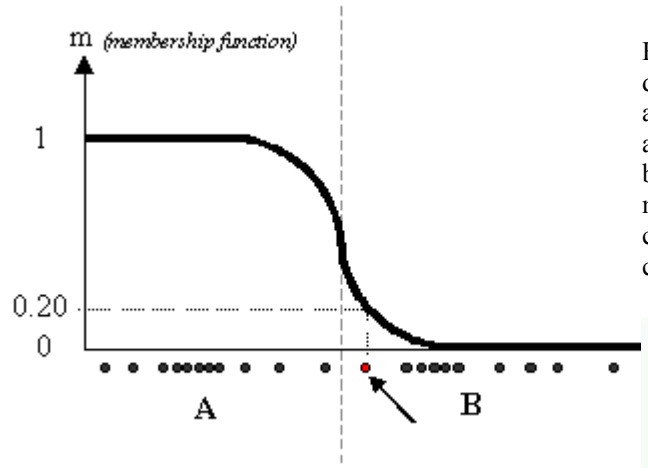
As already told, data are bound to each cluster by means of a Membership Function, which represents the fuzzy behaviour of this algorithm. To do that, we simply have to build an appropriate matrix named U whose factors are numbers between 0 and 1, and represent the degree of membership between data and centers of clusters.
For a better understanding, we may consider this simple mono-dimensional example. Given a certain data set, suppose to represent it as distributed on an axis. The figure below shows this:



Looking at the picture, we may identify two clusters in proximity of the two data concentrations. We will refer to them using 'A' and 'B'. In the first approach shown in this tutorial - the k-means algorithm - we associated each datum to a specific centroid; therefore, this membership function looked like this:



In the FCM approach, instead, the same given datum does not belong exclusively to a well defined cluster, but it can be placed in a middle way. In this case, the membership function follows a smoother line to indicate that every datum may belong to several clusters with different values of the membership coefficient.



In the figure above, the datum shown as a red marked spot belongs more to the B cluster rather than the A cluster. The value 0.2 of 'm' indicates the degree of membership to A for such datum. Now, instead of using a graphical representation, we introduce a matrix U whose factors are the ones taken from the membership functions:

$$U_{MxC} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ .. & .. \\ 0 & 1 \end{bmatrix} \qquad U_{MxC} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .. & .. \\ 0.9 & 0.1 \end{bmatrix}$$

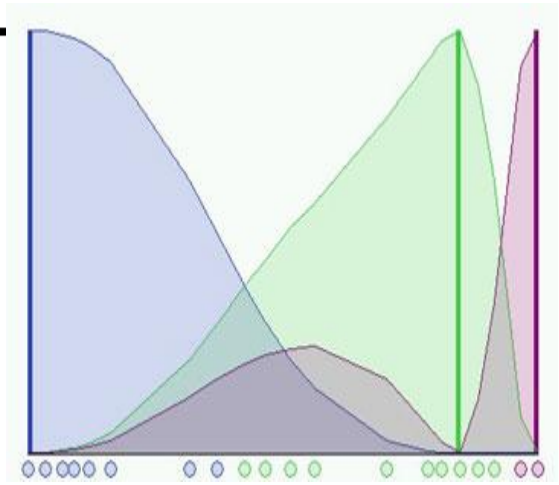
(a) (b)

The number of rows and columns depends on how many data and clusters we are considering. More exactly we have C = 2 columns (C = 2 clusters) and N rows, where C is the total number of clusters and N is the total number of data. The generic element is so indicated: u_{ij}.
In the examples above we have considered the k-means (a) and FCM (b) cases. We can notice that in the first case (a) the coefficients are always unitary. It is so to indicate the fact that each datum can belong only to one cluster. Other properties are shown below:

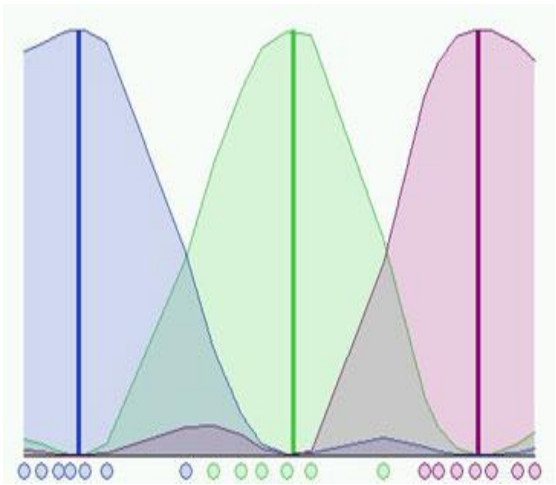
- $u_{ij} \in [0,1] \quad \forall i, j$
- $\sum_{j=1}^{C} u_{ik} = 1 \quad \forall i$
- $0 < \sum_{i=1}^{N} u_{ij} < N \quad \forall N$

An Example

Here, we consider the simple case of a mono-dimensional application of the FCM. Twenty data and three clusters are used to initialize the algorithm and to compute the U matrix. Figures below (taken from our interactive demo) show the membership value for each datum and for each cluster. The color of the data is that of the nearest cluster according to the membership function.



In the simulation shown in the figure above we have used a fuzzyness coefficient m = 2 and we have also imposed to terminate the algorithm when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < 0.3$. The picture shows the initial condition where the fuzzy distribution depends on the particular position of the clusters. No step is performed yet so that clusters are not identified very well. Now we can run the algorithm until the stop condition is verified. The figure below shows the final condition reached at the 8th step with m=2 and ε=0.3:



Is it possible to do better? Certainly, we could use an higher accuracy but we would have also to pay

for a bigger computational effort. In the next figure we can see a better result having used the same initial conditions and ε=0.01, but we needed 37 steps!

It is also important to notice that different initializations cause different evolutions of the algorithm. In fact it could converge to the same result but probably with a different number of iteration steps.

REFERENCES:

1. http://en.wikipedia.org/wiki/Computer_cluster
2. http://msdn.microsoft.com/en-us/library/ff649250.aspx
3. http://en.wikipedia.org/wiki/Cluster_analysis
4. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html
5. http://en.wikipedia.org/wiki/Fuzzy_clustering
6. http://www.cs.utexas.edu/users/csed/doc_consortium/DC99/wooley-abstract.html
7. http://www.slideshare.net/dataminingcontent/data-mining-clustering-and-analysis-5006093

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

WEB PERSONALIZATION ON TASK BASED MODELING IN WEB USAGE MINING

[1]Dr. A. Saradha, [2]Mrs. T. Mohana Priya,Research Scholar

[1]Head, Department of Computer Science, IRTT, Erode, Tamilnadu, India

[2]Bharathiar University, Coimbatore, Tamilnadu, India

ABSTRACT

Web Usage Mining(WUM) is the process of taking out interesting behavior patterns that allow analyzing uses by the website administrator. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URL's accessed. . The primary goal of the search engine is to provide the relevant information to the users to fulfill their needs. One approach to satisfy the requirements of the user is to personalize the information available on the web, called Web Personalization. We describe a method whereby we mine historic search-engine logs to find other users performing similar tasks to the current user and leverage their on-task behavior to identify Web pages to promote in the current ranking. Providing richer models of the web page to the current user from the historic users' search tasks can help to improve likelihood of finding relevant content and enhance the concept of personalization. In this paper presents elaborately how to identify the similar tasks to the current users' task from the web log files and ranking to the web pages.

Key Terms:

Web Usage Mining, Personalization, Web Log Files, Ranking Algorithm, Groupization

1. INTRODUCTION

The World wide Web is a rich source of information , expand in size and complexity. An ultimate need of the search engine is that of predicting the user needs in order to improve the usability of a web site. Web Personalization can be defined as any action that adapts the information or services provided by a web site to an individual user, or a set of users, based on knowledge acquired by their navigational behavior, recorded in the Web Log files. Historic search interactions from a user over time can be used to personalize search results [3,4], but the focus there is either once again on query based matching [4] or creating general models of searcher interests across a variety of topics [3].

This paper is structured as follows. Section 2 describes Web Mining and Web Mining Categories, section 3 describes Preprocessing Stages, Section 4 describes related work in task modeling, personalization, and miningtask-relevant search behavior. Sections 5 and 6 describe the identifying similar tasks and ranking features. Section 7 and 8 discusses these findings and implications and we conclude in Section 9.

2. WEB MINING:

Web Mining is the Data Mining technique that automatically discovers or extracts the information from web documents. There are three areas in web data mining.

2.1 Web Content Mining

It is the process of extracting the information from the content of the web pages. Web content mining is related to data mining techniques can be applied in web content mining. It is also related with text mining , because web data is mainly semi structured in nature.

2.2 Web Structure Mining

Web structure mining aims to generate structural summary about web sites and web pages. The

focus of structure mining is therefore on link information, which is an important aspect of web data.

2.3 Web Usage Mining

It is used to discover interesting usage patterns from web data. in order to understand and better serve the needs of web based application. It tries to make sense of data generated by the web surfer's sessions/ behaviors.

3. PREPROCESSING

Web log data is usually diverse and voluminous in nature. This data must be assembled into a consistent, integrated and comprehensive view. A typical example of web log file is shown in Figure.1.

<ip_addr><base_url> - <date><method><file><protocol><code><bytes><referrer><user_agent>

Figure 1 : Common Web Log Format

A web server log file contains requests made to the webserver, recorded in chronological order. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. A common log format file is created by the web server to keep track of the requests that occur on a web site. A standard log file has the following format as shown in Figure 2.

152.152.98.11 - -[16/Nov/2011:16:32:50 -0500] "GET /jobs/ HTTP/1.1"200 15140 "http://www.google.com/search?q=salary+for+data+mining&hl=en&lr=&start=10&sa=N" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)"

Figure 2: Example of Log File format

A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. When user submit request to a web server that activity are recorded in web log file. Log file range 1KB to 100MB.

```
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:21 -0600] "GET /Calls/ONOM.html
HTTP/1.0" 200 3942 "http://www.lycos.com/cgi-
bin/pursuit?query=advertising+psychology&maxhits=20&cat=dir" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:23 -0600] "GET
/Calls/Images/earthani.gif HTTP/1.0" 200 10689 "http://www.acr-news.org/Calls/ONOM.html"
"Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:24 -0600] "GET /Calls/Images/line.gif
HTTP/1.0" 200 190 "http://www.acr-news.org/Calls/ONOM.html" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:25 -0600] "GET /Calls/Images/red.gif
HTTP/1.0" 200 104 "http://www.acr-news.org/Calls/ONOM.html" "Mozilla/4.5 [en] (Win98; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:31 -0600] "GET / HTTP/1.0" 200 4980 ""
"Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/line.gif
HTTP/1.0" 200 190 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/red.gif
HTTP/1.0" 200 104 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/earthani.gif
HTTP/1.0" 200 10689 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:33:11 -0600] "GET /CP.html HTTP/1.0" 200
3218 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"
```

Figure 3 : Example of typical server log

A CLF file is created by the Web server to keep track of the requests that occur on a Web site. Figure 3 is shown the typical server log. The stages of preprocessing are shown in Figure.4

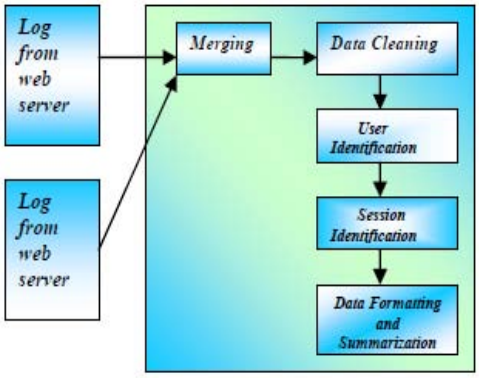


Figure 4 : Stages of Preprocessing

It comprises of the following steps- Merging of log files from Different Web Servers, Data Cleaning, Identification of users, Sessions, and Visits, Data Formatting and Summarization.

3.1.Merging

At the beginning of the data preprocessing, the requests from all log files in Log, with the web server name to distinguish between requests made to different Web Servers and taking into account the synchronization of web server clocks, including time zone differences. The merging problem is formulated as " i en the set of log files Log={L_1, L_2, L_3….L_n }, merge these log files into single log file £(joint log file)".

3.2.Data Cleaning

The second step of data preprocessing consists of removing useless request from the log files. For

example, requests for graphical page content(*.jpg &*.gif images) and requests for any other file which might be included into a web page or even navigation sessions performed by robots and web spiders.

3.3. User Identification

In most cases, the log file provides only the computer address and the user agent. For websites requiring user registration, the log file also contains the user login that can be used for the user identification. When the user login is not available, each IP is considered as a user.

3.4. Session Identification

A user session is a directed list of page accesses performed by an individual user during a visit in a Web site. A user may have a single(or multiple) session(s) during a period of time.

3.5. Data Formatting & Summarization

Here the structured file containing sessions and visits are transformed to a relational database model. Then, the data generalization method is applied at the request level and aggregated for visits and user sessions to completely fill in the database. The data summarization concerns with the computation of aggregated variables at different abstraction levels(e.g request, visit and user session).

4. RELATED WORK

There are three relevant areas of related work: (1) task modeling (2) personalization of search engines(3) mining the search behavior of other users.Monika Shoni, Rahul Sharma et al.[1] provides framework for web personalization using web mining. K.R. Suneetha, et al. [2] focuses on the analysis of website top errors, potential visitors of the site, etc., Fang Yuan, et al. [5] mainly focus on analyzing visiting information from logged data in order to extract usage paterns, which can be classified on to three categories. Rekha Jain, et al. [6] discusses the page ranking algorithm for web mining. Ramya C, et al.[7] discusses stages of preprocessing and Web log files.

5. IDENTIFYING SIMILARTASKS

The first step is applying task identification in sessions. Before that , we have to use groupization concept to collect historic users' search tasks. To identifying similar tasks involves two process. There are Groupization and Computing task Similarity.

5.1 Groupization

The task-based approach can be applied to the current user's search history, or as we focus on here, all users' search histories as so-called "groupization" (a variant of personalization hereby other users' profiles can be used to personalize the search experience[16]).The first step in applying our method is to identify tasks within search sessions. Now we describe the task identification Process. It involves in two steps: 1) Log data 2) Identifying task in sessions.

5.1.1 Log Data

The primary source of this study is a data set comprising from the anonymized set of users of the Google Search Engine. The logs contained a unique user identifier, a search sessionidentifier, the query, the top-10 URLs returned by the search engine for that query, and clicks on the results.Logs were split into search sessions demarcated with a 30-minute inactivity timeout, such as that used in previous work[17].

5.1.2 Identifying tasks in sessions

In order to calculate inter-query similarities, QTC takes a supervised learning approach.It involves two process.

a) Measure the similarity between query pairs
b) Cluster queries into tasks based on similarities

Figure 5 explains how the sessions and tasks are divided. Here, with the help of query clustering QTC[8], which has the advantage of segmenting the interleaved tasks in a session.

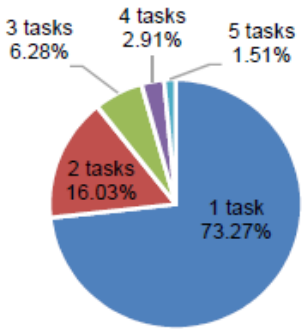


Figure 5 : Number of search tasks in search sessions

Figure 5 illustrates the fraction of sessions containing between one and five search tasks. The figure shows that around 90% of sessions have one or two tasks; 73.3% sessions contain a single task and about 16.0% sessions contain two tasks. This

shows that although most sessions comprise a single task, there are still a sizable number of sessions (over 25%) containing multiple tasks.

5.2 Computing Task Similarity

A key part of this process is finding other users attempting searching similar tasks. In this section describes how find similar task and features that we generate for ranking.There are number of ways to find similarity between a given pair of tasks.

5.2.1 Query Similarity

These similarity measures are based on comparing the queries thatusers issue in both tasks under consideration. Similarity in this casecan be based on the exact terminology used in the queries (afternormalization) and more generally, on the semantic similarity betweenthe queries.

5.2.2 Syntactic Similarity and Semantic Similarity

Syntactic similarity describes the string match between the queries.Similarity can be computed based on the overlap between the tasksin terms of:

(1) the fraction of queries that are shared between tasks(i.e., the intersection divided by the union), and

(2) as the fractionof unique query terms that are shared between tasks.

While the queries may not overlap, but the semantic of queries may overlap. To address this, compute task similarity by measuring semantic similarity. Let Q = $q_{1.....}q_j$be one query and S =$s_{1....}s_I$ *be* another, the semantic similarity between these two queries canbe measured based on the IBM Model 1 [9,10]. Treating Qand Sas twosequences of words, the IBM Model 1-based semantic similarity model is defined as:

$$P(S|Q) = \prod_{i=1}^{I} \sum_{j=1}^{J} P(s_i|q_j) P(q_j|Q)$$

Where P(q|Q) is the unigram probability of word qin query Q. The word translation probabilities P(s|q)are estimated on the query title pairs derived from the clickthrough search logs, assuming that the title terms are likely to be the desired alternation of the paired query.

5.2.3 Clicked-Result Similarity

The result URLs that are clicked (and have an associated longdwell) provide a different source of information about users' searchintent than is available in queries. We compute similarity in threeways: (1) the match between the clicked URLs, (2) the match betweenthe domains of the clicked URLs, and even more generally,(3) the match between the topical categories assigned to the URLs.

5.2.4 URL and Domain Similarity

Similarity is computed based on the overlap between the tasks interms of the fraction of unique clicked URLs shared between them(i.e., the intersection of clicked URLs divided by the union of allclicked URLs), or in terms of the URL domains that are shared betweenthe tasks. Backing off to Web domains provides more opportunityfor a match between clicked results, improving coveragewhile preserving information about the web site of interest.

5.2.5 Topical Similarity

Rather than relying on exact matches between particular URLs ortheir domains, we can also consider matching based on the topicalityof the pages described earlier. Given a categorization for eachof the clicked URLs for a task t, for that task, with a probability for each topical categoryc(i.e., $P_t(c)$). Given this representation, we can compare thedistribution for the current task with the distribution over all othertasks from other users {t'}. We perform this comparison using bothKullback-Liebler divergence and the cosine similarity. That is,

$$KL(t', t) = \sum_{c \in C} \ln\left(\frac{P_{t'}(c)}{P_t(c)}\right) P_{t'}(c)$$

$$cos(C_{t'}, C_t) = \frac{C_t \cdot C_{t'}}{\|C_t\| \, \|C_{t'}\|}$$

Using two measures focused on different aspects of the distributionalsimilarity between tasks: KL computes the information gainand is asymmetric, cosine similarity computes the normalized dotproduct between the distributions (as vectors) and is symmetric.

5.3 Mining Similar Tasks

In this section we describe the procedure that we employ toleverage task similarity in re-ranking the top retrieved results, aswell as the different groups from which we can find similar tasks.Foreach of the tasks t′ in the set of all tasks observed historically fromother usersT, we can then compute the similarity between the currenttask and those tasks k(t ,t').For each of the similaritymeasures in

Section 5.2. we can compute a score $s_k$for a URL uappearing in the top 10 search results :

$$s_k(t,u) = \sum_{t' \in T} \left(k(t,t') \cdot w(t',u)\right)$$

Where w(t',u), is a weight reflecting the importance of the URL ina related task. In our case we define w(t',u)as the click frequencyon that URL for the related task. Table 1 shown the possible similarities between task t and t'.

Table 1. Features used to compute the similarity between two tasks, t and t', in the computation of $k(t,t')$.

Feature name	Definition
FullQueryOverlap	The fraction of all queries in the union of t and t' that the two tasks have in common.
QueryTermOverlap	The fraction of all unique query terms in the union of t and t' that the two tasks have in common.
QueryTranslation	Semantic similarity between the queries in t and the queries in t' ($P(E\|Q,\theta)$ as defined earlier).
ClickedURLOverlap	The fraction of clicked URLs in the union of t and t' that the two tasks have in common.
ClickedDomainOverlap	The fraction of clicked domains in the union of t and t' that the two tasks have in common.
CategorySimilarityKL	The Kullback-Liebler divergence between the ODP category distribution from result clicks in t versus the same distribution from t'.
CategorySimilarityCosine	The cosine similarity between the ODP category distribution from result clicks in t versus the same distribution from t'.

6. RANKING FEATURES

Web mining technique provides the additional information through hyperlinks where different documents are connected. There are number of algorithms proposed for ranking to the web pages.

6.1 PageRank

This algorithm was developed by Brin and Page at Stanford University[11]. The Page Rank forms a probability distribution over the web pages so the sum of Page Ranks of all web pages will be one. Page and Brin proposed a formula to calculate the PageRank of a page A stated as below

PR(A) = (1-d)+d(PR(T_1)/C(T_1) + ……..+PR(T_n/C(T_n))

Here PR(T_i) is the PageRank of the Pages T_i which links to page A, C(T_i) is the number of outlinks on page T_i and d is damping factor. It is used to stop other pages having too much influence.

6.2 Weighted Page Rank

This algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an extension of PageRank algorithm[12]. The importance is assigned in terms of weight values to incoming and outgoing links.This is denoted as $W^{in}_{(m,n)}$ and $W^{out}_{(m,n)}$ respectively.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \in R(m)} I_p}$$

In is number of incoming links of page n, Ip is number of incoming links of page p, R(m) is the reference page list of page m.

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p \in R(m)} O_p}$$

On is number of outgoing links of page n, Op is number of outgoing links of page p, Then the weighted PageRank is given by formula in

$$WPR(n) = (1-d) + d \sum_{m \in B(n)} WPR(m) W^{in}_{(m,n)} W^{out}_{(m,n)}$$

6.3 ReRanking Feature

The inter-task similarity is computed using k(t ,t') which isthen multiplied against the click count for each URL in the top-10(i.e., w(t',u)). The result summed over all tasksin the historic data is used to generate the final feature value.In addition, we also compute *ClickedTasksCount*, which is the totalnumber of tasks for which a particular URL uis clicked. Thismeasures URL popularity independent of task. Note that since *QueryTranslation* and *CategorySimilarityKL* are asymmetric.

7. EXPERIMENTS

Our log-based evaluation method focuses on a re-ranking task, assessing the extent to which the models promote clicked results.The re-ranking models attempt to promote observed *satisfied resultclicks* (SAT clicks) toward higher rank positions in the result list. This enables offline evaluation of models performance using judgments personalized to each user. This approach has been used to determine the effectiveness of various re-ranking methods [13,14,15].We define the clicks having less than 30 seconds dwelling time as *quickbacks*. We consider three types of clicks in labeling user feedback in the logs: SAT clicks, *quickback* clicks, and no clicks.

Table 2. Statistics of the weekly data for learning/evaluation.

Count	Training	Validation	Test
SAT Clicks	2,086,335	2,062,554	2,082,145
Quickback Clicks	417,432	408,196	413,496
Tasks	1,165,083	1,126,452	1,135,320
Queries per Task	1.678	1.676	1.666

In each impression, if a URL received atleast one SAT click, the URL is labeled with a 2; if a URL receivedonly quickback clicks, the URL is labeled with a 1; if a URL wasnot clicked at all, the URL is labeled with a 0. This gives us a threeleveljudgment for each top-10 URL for each query.

7.1 Measures

We measure ranking quality by mean average precision and meanreciprocal rank. In both cases, the mean is calculated over all theimpressions in our test set. Mean average precision (MAP) for a setof queries is the mean of the average precision scores for eachquery. The average precision score is defined as

$$AveragePrecision = \frac{\sum_{k=1}^{n} Precision(k)Rel(k)}{\sum_{k=1}^{n} Rel(k)}$$

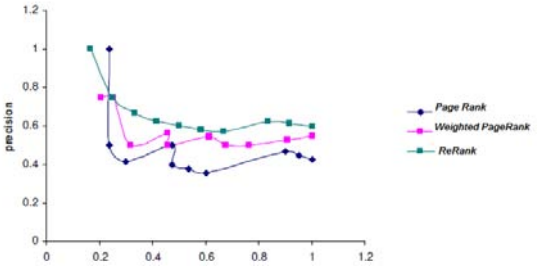
where nis the number of URLs in the impression, usually 10, Rel(k)is an indicator function equaling 1 if the URL at rank is relevant document, zero otherwise, and Precision(k)is the precisionat cut-off kin the ranked list. Mean reciprocal rank (MRR) for a query set is the average of thereciprocal ranks across all results, which is defined as

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$$

Where $rank_i$ is the rank of the first relevant URL in the ranking list,and Nis the number of impressions in test.These measures are complementary in that MRR focuses on therank of the first relevant document in the top 10, whereas MAP targetsthe rank of relevant results across the top 10 documents.

8. RESULT

Precision-Queries graph is done on average of user feedbacks. The graph shows that proposed ReRank gives better results than PageRank and Weighted PageRank. We managed to provide high precision at the tasks increases.



Queries

Figure 6 : Precision – Queries Plot

9. CONCLUSION

This work clearly demonstrates the value of considering search *tasks* rather than just search *queries* during personalization, as well as the benefit ofgroupization.In this chapter we have attempted to present a comprehensive view of the personalization process based on web usage mining. The approaches we have detailed show web log files, identifying similar tasks, Re ranking features, can be leveraged effectively as an integrated part of a web personalization system. In future work involves use of a broader range of cohorts. Cohorts include location, browser, place, etc., and the development of more sophisticated and generalizable models of task behavior.

REFERENCES

[1] Monika shoni,Rahul Sharma and Vishal Shrivatsava. Framework for web Personalization. IJRET Oct 2012

[2] K.R. Suneetha, Dr. R. Krishnamoorthy, "identifying User Behavior by analyzing web server Access Log Files", International Journal of Computer Science and Network Security, Apr 2009

[3] Sontag, D., Collins-Thompson, K., Bennett, P.N., White,R.W., Dumais, S.T., and Billerbeck, B. (2012). Probabilistic models for personalizing web search. *WSDM*, 433–442.

[4] Teevan, J., Liebling, D.J., and Geetha, G.R. (2011). Understanding and predicting personal navigation. *WSDM*, 85–94.

[5] Fang Yuan, Li- Juan Wang, Ge Yu, "study on data Preprocessing algorithm in Web Log Mining" Proceedings of the Second

International Conferences on Machine Learning and Cybernetics, Nov-2003.

[6] Rekha Jain, Dr. G.N> Purohit, "Page Ranking algorithm for Web Mining", International Journal of Computer Applications, January 2011.

[7] Ramya C, Shreedharan K.S, and Kavitha G, "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process", International Journal of Electronics Engineering, March 2013.

[8] Liao, Z., Song, Y., He, L.W., and Huang, Y. (2012). Evaluating the effectiveness of search task trails. *WWW*, 489–498.

[9] Berger, A.L. and Lafferty, J. (1999). Information retrieval asstatistical translation. *SIGIR*

[10] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., and Mercer,R.L. (1993). The mathematics of statistical machine translation:parameter estimation. *Comp. Ling.*, 19(2): 263–311.

[11] S.Brin and L.Page, " The Anatomy of a Large Scale Hypertextual Web Seaarch Engine", Computer Network and ISDN System , 1998.

[12] J. Kleinberg, Authoritative sources in a Hyper-linked Environment , Journal of the ACM,1999.

[13] Bennett, P.N., Radlinski, F., White, R.W., and Yilmaz, E.(2011). Inferring and using location metadata to personalized web search.

[14] 5. Bennett, P., White, R.W., Chu, W., Dumais, S., Bailey, P.,Borisyuk, F., and Cui, X. (2012). Modeling the impact of shortand long-term behavior on search personalization.

[15] Sontag, D., Collins-Thompson, K., Bennett, P.N., White,R.W., Dumais, S.T., and Billerbeck, B. (2012). Probabilisticmodels for personalizing web search.

[16] Ryen W. White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, Hongning Wang, "Enhancing personalized search by mining and modeling task behavior",2013.

[17] Downey, D., Dumais, S.T., Liebling, D., and Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Replication Attacks Detection protocol in Wireless Sensor Networks

L.DEVI., MCA.,MPhil.,

Associate professor in PG Computer Science, Muthayammal college of Arts & Science,Rasipuram, Namakkal(Dt)

ABSTRACT

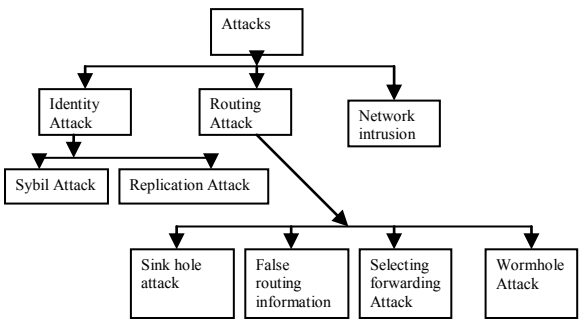
A wireless sensor network (WSN) consists of a number of tiny, low-cost, and resource-constrained sensor nodes, but is often deployed in unattended and harsh environments to perform various monitoring tasks. As a result, WSNs are susceptible to many application-dependent and application-independent attacks. Sensor networks deployed in hostile areas are subject to node replication attacks, in which an adversary compromises a few sensors, extracts the security keys, and clones them in a large number of replicas, which are introduced into the network to perform insider attacks. Memory overhead, energy efficiency and detection probability are the main technical concerns for any replication detection protocol. We propose four replication detection protocols that have high detection probability low memory requirement, and balanced energy consumption. The new protocols use Bloom filters to compress the information stored at the sensors, and use two new techniques, called *cell forwarding* and *cross forwarding*, to improve detection probability, further reduce memory consumption, and in the mean time distribute the memory and energy overhead evenly across the whole network. We classify existent detections in the literature, and explore the various proposals in each category. We look into necessary technical details and make certain comparisons, so as to demonstrate their respective contributions as well as limitations. We also present the technical challenges and indicate some possible directions for future research.

Introduction

The WSN is built of "nodes" – from a few to several hundreds or even thousands, where each node is connected to one (or sometimes several) sensors. Each such sensor network node has typically several parts: a radio transceiver with an internal antenna or connection to an external antenna, a microcontroller, an electronic circuit for interfacing with the sensors and an energy source, usually a battery or an embedded form of energy harvesting. Attack means any kind of malicious activity that attempts to collect, disrupt, deny, degrade, or destroy information system resources or the information itself.

We classify the network attacks into three main categories. It will show in the figure. The identity attacks are Sybil attack and clone (Replication)attack. In a Sybil attack the WSN is subverted by a large number of fake identities in order to disrupt the network's protocols. A node replication attack is an attempt by the adversary to add one or more nodes to the network that use the same ID as another node in the network.

We consider the node replication attack, which is an application-independent attack unique to wireless sensor networks. The attack makes it possible for an adversary to prepare her own low-cost sensor nodes and induce the network to accept them as legitimate ones. To do so, the adversary only needs to physically capture one node, reveal its secret credentials, replicate the node in large quantity, and deploy these malicious nodes back into the network so as to subvert the network with little effort. Recently, Ko et al. proposed a neighbor-based detection scheme to cope with replication attacks. The scheme features distributed detection and takes node mobility into account. It harnesses the dynamic observations of the neighbors of a claimer node and avoids the protocol iterations typically found in distributed detections. Unfortunately, we show that their proposal is subject to various replication attacks that can circumvent the detection. Moreover, it is even possible for a sophisticated adversary to exploit the protocol to revoke legitimate nodes.



In this comprehensive survey, we consider a very severe and important physical attack on WSN which is called node replication attack or clone attack. It is also known as identity attack. In this attack, an adversary first physically captures only one or few of legitimate nodes, then clones or replicates them fabricating those replicas having the same identity (ID) with the captured node, and finally deploys a capricious number of clones throughout the network. This whole process of node replication attack and the various stages



This vexing problem arises from the actuality that sensor nodes are unshielded. It is stated in [2] that an experienced attacker can completely compromise a typical sensor node by using only a few readily available tools, and it can then obtain copies of that node memory and data within 1 min of discovering it. The clones or replicas may even be selectively reprogrammed to subvert the network by launching further insider attacks like falsifying sensor data or suppressing legitimate data, extracting data from the network and disconnect the network by triggering correct execution of node revocation protocols that rely on threshold voting schemes and staging denial of service (DoS) attacks. Clone nodes may create a black hole, initiate a wormhole attack with a collaborating adversary, or may also leak data in an environment in which sensed data must be kept private [3]. If these replicated nodes or clones remain undetected or unattended for a long time, they can further commence the changes in protocol behavior and intrusion into the systems security [4]. It is easy for an adversary to launch such attacks due to the fact that the clones, created by an adversary, have legitimate information (codes, key materials, and credentials), and they may be considered as legitimate nodes and totally honest by its neighbors which are participating in the network operation in the same way as the noncompromised nodes.

The above mentioned traditional security schemes for WSNs are inept to detect and prevent node replication attack. Thus, in the last few years, a number of detection and prevention techniques/schemes have been proposed in the literature. According to [7], the detection schemes are classified on a high level as network-based or radio-based detection. Only one instance of radio-based detection is found in [6]. The former category is further categorized into two types as for mobile WSNs and for stationary WSNs. Both techniques for mobile and stationary WSNs are further divided into two broad categories, namely, centralized and distributed.

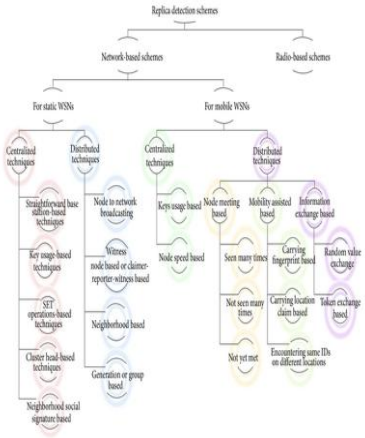


Figure 3 which shows a detailed classification of all replica detection schemes. This categorization provides a first step to better understand the node replication detection schemes

Memory Efficient Protocols for Detecting Node Replication Attacks in Wireless Sensor Networks

These techniques fall into the category of witness node-based techniques. Zhang et al. [1] have proposed four memory efficient multicast protocols for replication detection, namely, memory efficient multicast with Bloom filters (B-MEM), memory efficient multicast with Bloom filters and Cell Forwarding (BC-MEM), memory efficient multicast with cross forwarding (C-MEM), and memory efficient multicast with cross and cell forwarding (CC-MEM). The first protocol B-MEM use Bloom filters to compress the information stored at the sensors and the location claim C of a node is multicast via its neighbors to a number of randomly selected locations in the network. Each neighbor has a probability to participate in the multicast. If it does, it becomes a witness node and sends C to a random location in the network. The node closest to that location will be another witness node w to store C. The watcher nodes on the routing path from to w only store the membership of ID and l in the Bloom filters. Such membership information can help them detect any conflicting location claim C' received later, and guide C' along to either or , which will then broadcast both C and C' to the entire network in order to revoke node and its replicas.

The second protocol BC-MEM is designed on top of B-MEM. It adopts a cell forwarding technique that not only solves the crossover problem

but also reduces the memory overhead. The deployment area is divided into virtual cells. In each cell an anchor point is assigned for every node in the network. The anchor point for a node is determined by ID. The node closest to the anchor point is called the anchor node for . In B-MEM, when a location claim is forwarded on a line segment, all intermediate nodes on the line serve as watchers, while the first node and the last node serve as witnesses. In contrast, in BC-MEM a claim is not forwarded on the line segment. It is forwarded to the anchor point in the next cell where the line segment intersects. The claim is forwarded from one anchor node to another until reaching the last cell. The anchor nodes in the intermediate cells are watchers, and the anchor nodes in the first and last cells are witnesses.

The third protocol C-MEM is designed on top of B-MEM. It incorporates a new cross forwarding technique to solve the crowded center problem. B-MEM stores the information about a location claim along randomly selected line segments, which are likely to pass the center area of the deployment. On the other hand, C-MEM first selects a random point (called the cross point) in the network and forwards the location claim to that point. From there, it forwards the claim along the horizontal and vertical lines that pass the cross point. While the node closest to the cross point is a witness node, the nodes along the horizontal and vertical lines are watchers. Since the cross points for all location claims are distributed uniformly at random in the network, it is no longer true that the lines pass the center area more frequently. C-MEM does not use cell forwarding.

The fourth protocol CC-MEM combines cross forwarding and cell forwarding to solve both the crowded center problem and the crossover problem, such that it can detect node replication attack with high probability and low overhead.

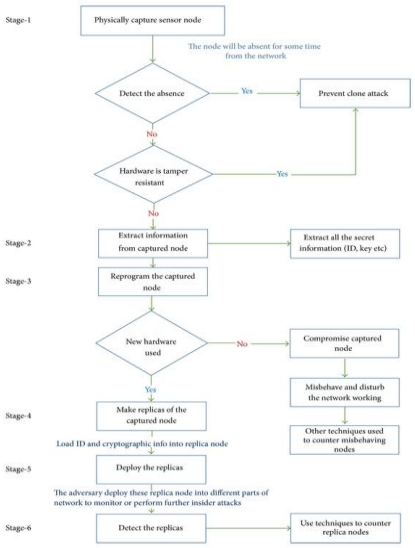
Active Detection of Node Replication Attacks

This technique falls into the category of base station-based techniques. Melchor et al. [5] have proposed a distributed protocol for the detection of replication attack for wireless sensor networks, in which each node verifies at random a few other nodes in the network. The proposed protocol does not build a distributed database of location claims that will contain local conflicting claims when replicas exist. The idea is that each node will actively test if 1k other random nodes are replicated or not; they call them the scrutinized nodes. In order to test whether a scrutinized node α is replicated or not, 2k nodes are randomly chosen in the network and asked to forward to α a request for a signed location claim. If two replicas exist, each will probably receive a request, and if both answer, two conflicting claims will be obtained by the queerer**.**

Discussion

Node replication attack or clone attack is one of the most harmful and dangerous threat to an unattended wireless sensor network because in this attack an adversary not only compromises the sensor nodes but can also carry out a large class of internal attacks for instance DoS attack, Sybil attack, and Black hole, and wormhole attack, by surreptitiously inserting arbitrary number of replicas at strategic positions of the network. First, since a clone or replica is considered to be completely honest by its neighbors, the legitimate nodes cannot be aware of the fact that they have a clone among them. Voting mechanisms remain unsuccessful to detect clone nodes that are not within the same neighborhood as a voting mechanism is used to detect misbehaving nodes and clones within the neighborhood to agree on the legitimacy of a given node. Thus, there is a need for global countermeasure that can detect clones on the global level. Second, the general purpose security protocols for secure sensor network communication would allow replica nodes to create pair-wise shared keys with other nodes and the base station, and thus in doing so, the replica nodes are able to encrypt, decrypt, and authenticate all of their communications as if they were the original captured nodes.

The process or stages of node replication attack can be described in the form of a flow chart as shown in Figure 5. Finally these replicas or clones can be detected by using various detection schemes



Since clone nodes carry all the cryptographic and keying materials, all the traditional authentication and intrusion detection techniques are ineffective to discover and detect these clones or replicas in the

network. Keeping this in mind many techniques have been proposed for the detection of node replication attack and recall that these are broadly categorized into centralized and distributed techniques.. One important class of distributed techniques is witness node-based techniques which are considered to the most favorable techniques yet for detecting clone nodes. But according to Zeng et al., replica detection protocols must be non-deterministic and fully distributed in order to circumvent the existing drawbacks of witness-based strategies. The witness node-based strategies ought to fulfill three requirements to have a high probability of detecting clones or replicas. Firstly, the selection of witness-nodes should be nondeterministic as it is more difficult for an adversary to launch clone attacks in nondeterministic protocols successfully because the witnesses of node are not known and are different in each execution of the protocol. Secondly, for any given node, all the nodes should have an equal probability to be the witnesses of that node during the lifetime of the network. Thirdly, the witness-nodes should be selected from all over the network randomly and not from particular area of the network every time meaning that the witness distribution should be uniform throughout the entire network.

There are two types of attacks which are the variations of node replication attack and can be launched by an adversary against witness node-based schemes. These are named as smart attack and masked replication attack. Smart attack is a special witness compromising attack, and in this attack an adversary avariciously chooses which sensor to corrupt in order to maximize its chance for its replicas to go undetected. The adversary finds out the witness nodes which are used to detect replicas and only compromises these witness nodes to avoid detection. The witness node-based techniques use a framework called claimer-reporter-witness framework in which a node referred to as claimer, locally broadcasts it location claim to its neighbors. Each neighbor serves as a reporter and employs a function to map the claimer ID to a witness. The neighbor forwards the claim to the witness and if it receives two different location claims for the same node ID then it means that the adversary has replicated a node. The adversary can also employ a function to know about the witness for the given claimer ID, and may also locate and compromise the witness node before she inserts the replicas into the wireless sensor network in order to evade the detection. In masked replication attack, the adversary may turn to compromise all the neighbors of a replica so as to prevent a location claim from propagating to any witness thus eliminating the reporters at all. This attack makes it possible for such a replica, whose neighbors have all been compromised, to lie about its physical position. So far, all the witness node-based techniques have assumed a static WSN, and are seemed to be the most promising schemes till yet to detect replicas or clones in static WSN, but alas these witness node-based schemes and location-based replication detection schemes are unable to detect and counter these types of replication attacks.The data accumulated in their memories become targets of many adversaries. In [11], a mobile adversary model is proposed in which mobile adversary visits and travels around the network trying to compromise a subset of sensors within the time interval when sinks are not present in the network. The time taken by a mobile adversary to compromise a set of sensors is much shorter than the time between two successive data collections of a sink. Thus, it is much difficult to snatch mobile compromised nodes as well as mobile clones.

Another challenge arises in mobile WSNs when a mobile adversary adopts a more sophisticated strategy named “group mobility strategy.” In this stratagem, the replicas form a physically close group which always moves together, but only a representative of them communicates with the genuine nodes, whereas the rest of the replicas remain inactive as “silent learners” so that they can learn (from encounters with genuine nodes) about any received token or corresponding meeting instant. Once the replicas have met all the “” genuine nodes (and thus acquired all the necessary knowledge to pass later authentications), they can scatter in the sensing region, and each behaves actively and independently, until the next detection round starts.

Also, when the mobile replicas communicate and collaborate with each other and share their keys or random numbers, they can make the detection technique fails to thwart them easily. Thus, mobile WSNs offer much more challenges in detecting mobile replicas, and it is highly needed to overcome these challenges by developing some new, different and more efficient detection techniques for detecting mobile replicas or clones.

Conclusion and future work

To improve detection probability further reduce memory consumption and in the mean time distribute the memory and energy overhead evenly across the whole network using simulation method we can complete. This paper reviewed the state-of-the-art schemes for detection of node replication attack also called clone attack. The existing techniques are broadly categorized into two classes distributed and centralized. Both classes of schemes are proficient in detecting and preventing clone attacks, but both schemes also have some noteworthy drawbacks. However, to sum up, the current study highlights the fact that there are still a lot of challenges and issues in clone detection schemes that

need to be resolved to become more applicable to real-life situations and also to become accepted by the resource constrained sensor node.

References

1. M. Zhang, V. Khanapure, S. Chen, and X. Xiao, "Memory efficient protocols for detecting node replication attacks in wireless sensor networks," in Proceedings of the 17th IEEE International Conference on Network Protocols (ICNP '09), pp. 284–293, Princeton, NJ, USA, October 2009.
2. C. Hartung, J. Balasalle, and R. Han, "Node compromise in sensor networks: the need for secure systems," Tech. Rep. CU-CS-988-04, Department of Computer Science, University of Colorado at Boulder, 2004.
3. H. Choi, S. Zhu, and T. F. L. Porta, "SET: detecting node clones in sensor networks," inProceedings of the 3rd International Conference on Security and Privacy in Communication Networks (SecureComm '07), pp. 341–350, September 2007.
4. S. Gautam Thakur, "CINORA: cell based identification of node replication attack in wireless sensor networks," in Proceedings of the IEEE International Conference on Communications Systems (ICCS '08), 2008.
5. C. A. Melchor, B. Ait-Salem, P. Gaborit, and k. Tamine, "Active detection of node replication attacks," International Journal of Computer Science and Network Security, vol. 9, no. 2, pp. 13–21, 2009.
6. S. Hussain and M. S. Rahman, "Using received signal strength indicator to detect node replacement and replication attacks in wireless sensor networks," in Data Mining, Intrusion Detection, Information Security and Assurance, and Data Networks Security 2009, vol. 7344 ofProceedings of SPIE, April 2009.
7. W. T. Zhu, J. Zhou, R. H. Deng, and F. Bao, "Detecting node replication attacks in wireless sensor networks: a survey," Journal of Network and Computer Applications, vol. 35, no. 3, pp. 1022–1034, 2012.
8. J. Deng, R. Han, and S. Mishra, "Security support for in network processing in wireless sensor networks," in Proceedings of the ACM Workshop on Security in Ad Hoc and Sensor Networks (SASN '03), pp. 83–93, 2003.
9. Y. Zeng, J. Cao, S. Zhang, S. Guo, and L. Xie, "Random walk based approach to detect clone attacks in wireless sensor networks," IEEE Journal on Selected Areas in Communications, vol. 28, no. 5, pp. 677–691, 2010.
10. I. Amundson and X. D. Koutsoukos, "A survey on localization for mobile wireless sensor networks," in Proceedings of the 2nd International Conference on Mobile Entity Localization and Tracking in GPS-Less Environments (MELT '09), vol. 5801 of Lecture Notes in Computer Science, pp. 235–254, 2009.
11. R. D. Pietro, L. V. Mancini, C. Soriente, A. Spognardi, and G. Tsudik, "Catch me (If you can): data survival in unattended sensor networks," in Proceedings of the 6th Annual IEEE International Conference on Pervasive Computing and Communications (PerCom '08), pp. 185–194, March 2008.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications
(NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Energy Efficient of Connectivity and Coverage Preserving Schemes for Surveillance Applications in WSN

[1]S.Meena [2] L.Devi.MCA.,M.Phil.,

[1]Research scholar, [2] Professor,
[1 & 2] Dept of computer science Muthayammal college of arts and science
Rasipuram(Tk),Namakkal(DT)

ABSTRACT

Wireless Sensor Networks(WSNs) are designed by sensor nodes that communicates each other and also processing data and sensing environment wirelessly[1]. The sensor nodes are widely used in various applications these days e.g. disaster management, border protection, battlefield surveillance, space exploration, secure installation etc. Sensor nodes are deployed to operate autonomously in unattended environments in these applications [2]. The figure 1 shows architecture of battlefield surveillance where each sensor collects the data to a base station.

1. Introduction

The networks performance can directly be affected by deployment, which is an important issue of WSNs [3]. The basic issue in WSN is the coverage problem and this problem is considered as a quality of service which can be provided by whole sensor network or by a single sensor node [4]. The main approach is to maintain the connectivity with the coverage fully preserved. The information coverage is defined according to sensor density for complete coverage [5].

In this paper, we are mainly focusing on the efficient coverage and the connectivity for randomly placement of the sensor nodes. To maintain the network connect or to maintain connectivity, we are going to use the connecting dominating set(CDS) as a virtual backbone of the network.

Deployment

The deployment scheme's choice is basically depends on the type of sensors, application and the environment. Nodes can be placed in an area of interest either deterministically or randomly[2]. Deployment of sensors nodes in wireless sensor network is an important issue as the performance of wireless sensor networks largely depends on deployment of sensors nodes. The deployment of sensor nodes is divided into two fractions according to the function of networks and these fractions are Coverage and Connectivity. The coverage is classified into static and dynamic category and the connectivity is divided into P-connectivity and RAB-connectivity [3].

Techniques for CDS construction

There are different techniques by which we can design CDS,among them some are following which are efficiently applicable to WSN.

Guha and Khuller's Algorithm

There are to two Greedy Heuristic algorithm's with bounded performance guarantee. Following are the strategies for building CDS.

CDS is grown from one node outward. Initially mark all nodes as white and select one white node which has maximum number of white neighbours.Mark that node as black remaining neighboring nodes as gray, now again select a node from gray nodes search for all white nodes, the one has maximum will be selected. Convert that gray node into black nodes and those white node into gray nodes. Iterate until all white nodes are covered.

WCDS is constructed and then select the intermediate nodes to create CDS.First color all nodes as white. Select a node which has maximum number of white neighboring nodes, color that node as black remaining as gray. Again select a white node having second maximum white neighboring nodes do the same exercise. After all connect those black nodes by coloring intermediate nodes as black. Make a Steiner Tree.

Connectivity preserving localized coverage algorithm for area monitoring WSN

Proposed solution Covers the area of interest(AOI),minimizing the count of active sensor nodes. If the overlap region of two sensor nodes is maximized the coverage is also maximized in region. It indicates 95% of coverage area and consuming 9.44J per unit in network. There is cluster head which process the nodes location and assign each sensor a set number under which the nodes are accessible. Once set region are defined later connection between sets established. Well this algorithm is deployed over large scale sensor nodes[12].

Connectivity based skeleton –Extraction in WSN

It is totally based on geometrical environment where the sensor nodes are deployed. In this Connectivity Based Skeleton Extraction(CASE) algorithm to compute skeleton graph. The graphs are generated by partitioning the boundary of Sensor network identifying the skeleton points, then from those points generate skeleton arcs. By these arcs a graph is generated for WSN[13].

Construction a CDS based Network backbone for energy efficiency in industrial WSN

In this we design a virtual backbone having k-way for CDS called as k- connected m-dominated(km CDS).Guha's and khuller's algorithm is implemented to design CDS.The main objective is to recover from fault tolerance, if a particular way is crushed we can go for other k-1 ways[13].

Coverage

The coverage problem in sensor networks is a basic issue, and considered to be a measure of the quality of service which can provided by a single sensor or by the entire network. The important aspect of coverage is to meet quality requirements, when sensors are deployed in the physical environment. If the amount of sensors deployed is less than the necessary requirement, then blind area may emerge and therefore lead to inaccurate tracking or monitoring of the area covered by the sensor network[4].The maximum coverage of some nodes in sensor networks can be described by an independent set of nodes, in which there are no intersections of sensing area between pair of nodes, an independent set of nodes is a set of nodes in graph, such that there is no link between any pair of nodes[5].The coverage issues are categorized into two parts i.e. Static and Dynamic[3].Certain tasks are performed by many WSN applications whose efficiency can be measured in terms of coverage, For these applications, it is necessary to define precise measures of coverage which impact overall system performance[6].For some WSN applications, surveillance of particular geographical areas is the main task. The coverage is basically related to the concept of area and node coverage. The parameter of area coverage is calculated by dividing the covered area to the total area of the terrain and the node coverage should be calculated by dividing number of redundant nodes to the total number of sensor nodes[7].

Connectivity

The key issue in sensor networks is to maintain the connectivity of the network to achieve a particular network performance. Each and every node in a network has a radio range, some nodes in a network can connect in a multi-hop manner, but not directly, increasing the radio range increase the connectivity and reduce network lifetime due to more transmission of power[3].It is must for the networks that deployment ensures that any pair of the nodes in the networks can reach each other, this is known as pure-connectivity (P-connectivity), which is independent of the employed routing algorithm, here power transmission is not considered and by considering power and performance, a specific routing algorithms can be used to make connection between arbitrary pair, is called routing algorithm based connectivity (RAB-connectivity)[3]. In [8], it has been shown that, for strong network connectivity, it is must to achieve complete coverage of area and the range of communication is at least twice of the sensing range.

Scheduling

The lifetime of network can be increased by scheduling only set of sensors that are necessary to be active and other sensors must to be sleep in a network, and the better lifetime can be achieved by reducing idle listening, traffic load etc [8]. In [9], many node scheduling methods are proposed and are divided into two main categories i.e.
round-based node scheduling and group-based node scheduling. According to round-based node scheduling, sensor nodes will execute the scheduling algorithm during the initialization of each round but according to group-based node scheduling, each node will execute the scheduling algorithm only once after its deployment and after the execution of scheduling algorithm, all nodes will be allocated into different groups.A connected dominating set CDS(G)⊆ of a graph G =(V,E), is defined as a subset CDS(G) V(G) of V(G) such that each node in V(G) – CDS(G) is adjacent to at least one node in CDS(G) and the produced graph (through CDS(G)) is a connected sub graph of G [10]. The nodes in sensor network communicate through multiple hops, and a connected CDS provides a backbone infrastructure [11]. In addition, backbone can be utilized for routing and activity scheduling.

2. Network MODEL

A WSN system is represented by a graph G(V,E)

comprised of a set of vertices V and edges E. The nodes u and v (u, v) ∈ V are said to be connected if both the nodes are within the communication range. The sensing range of node v ∈V is termed as coverage area monitored by the node.

Table 1: Observations

Number of nodes	Area	Percentage of nodes in backbone	No of nodes connected in backbone (nodes in backbone + nodes connectivity with backbone)	Percentage of nodes not connected to backbone
25	100*100	32%	20	20%
50	100*100	24%	48	4%
75	100*100	16%	71	5.3%
100	100*100	12%	88	12%
25	200*200	4%	7	72%
50	200*200	2%	8	84%
75	200*200	20%	33	56%
100	200*200	14%	40	60%

Conclusion

In this paper we try to find an efficient solution for WSN. As our algorithm having set, approach to select the communication region. The fascinating thing about our algorithm is the set making we activate a node in a set and sleep all other nodes which are in the same communication region because of this we can save the power by switching off all other nodes which are defined in the same set. Later on we are communicating between those sets. So it takes less energy to communicate, a region is defined and less number of nodes has to be in active mode. By statistical figure we can say that near about only 10% or 15% of nodes has been in active mode to make the communication efficiently. We had designed graphs by which we can conclude the regions of communication and how the data will move from one node to another node.

There is also a serious disadvantage of algorithm of isolated nodes. As there is possibility that some nodes might be in the region which is completely away from the range of Communication region, we can't establish any sort of communication between them, which is a serious impact of our algorithm.

References

[1] J. You, D. Lieckfeldt, J. Salzmann and D. Timmermann, GAF & Co : Connectivity Aware Topology Management for Sensor Networks, IEEE-2009

[2] M.Younis, K.Akkaya, Stratigies and techniques for node placement in wireless sensor networks:A Survey, Ad Hoc Networks,Vol. 6(June 2008).

[3] L.Liu, F.Xia , Z.Wang, J.Chen, Y.Sun, Deployment Issues in Wireless Sensor Networks , © Springer-Verlag Berlin Heidelberg 2005.

[4] Azzedine Boukerche and Xin Fei, A coverage-preserving scheme for wireless sensor network with irregular sensing range, Ad Hoc Networks Vol 5(2007).

[5] Wei An, F-M. Shao, H. Meng, The coverage-control
optimization in sensor network subject to sensing area ,Elsevier- 2008.

[6] A. Ghosh and S.K. Das, Coverage and connectivity issues in wireless sensor networks: A survey, Pervasive and Mobile Computing, Vol. 4, Elsevier-2008

[7] S. Oktug, A. Khalilov H. Tezcan, 3D Coverage Analysis under Heterogeneous Deployment Strategies in Wireless
Sensor Networks, Fourth Advanced International Conference on Telecommunications, IEEE, 2008.

[8] Q. Zhao, M. Gurusamy, Connected K-target coverage problem in wireless sensor networks with different observation scenarios, Computer Networks Vol.52,Elsevier-2008.

[9] L. Wang, R. Wei, Y. Lin, B. Wang, A clique base node scheduling method for wireless sensor networks,Journal of Network and Computer Applications Vol. 33, 2010.

[10] R. Misra, and C. Mandal, Minimum Connected Dominating Set Using a Collaborative Cover Heuristic for Ad Hoc Sensor Networks, IEEE Transactions on Parallel and Distributed Systems, vol. 21, no.3, March 2010.

[11] H. Raei, M. Tabibzadeh, B. Ahmadipoor, S. Saei, A Self - Stabilizing Distributed Algorithm for Mminimum Connected Dominating Sets in Wireless Sensor Networks With Different Transmission Ranges, ICACT 2009.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Lightweight Authentication Protocols for Low-Cost RFID Tags

[1]P.Nagarajan M.C.A.,[M.Phil(CS) [2]L.Devi M.C.A.,M.Phil.,PGDBA.,

[1] Research Scholor [2]Assistant Professor,
PGCS Department Muthayammal College of Arts & Science,
Rasipuram-637 408

ABSTRACT

Providing security in low-cost RFID tags is a challenging task because tags are highly resource constrained and cannot support strong cryptography. Special lightweight algorithms and protocols need to be designed that take into account the limitations of the tags. In this paper, we propose a set of extremely lightweight tag authentication protocols. We also provide an analysis of the proposed protocols.

1 INTRODUCTION

Radio Frequency Identification (RFID) systems are composed of RF tags and RF tag readers. Most tags consist of an antenna connected to a microchip. The use of silicon-based microchips enables a range of functionality to be integrated into the tags, including readable/writable storage and limited computing capability. Tag readers broadcast an RF signal to access information stored on the tags. This information can range from static identification numbers to user written data or data computed by the tag.

In the near future, low-cost RFID tags attached to consumer items as "smart-labels" may become an economical and efficient replacement for optical bar codes. Indeed, RFID tags offer several advantages over optical bar codes: data may be read automatically, without line of sight, through non-conducting material, at a rate of several hundred tags per second, and from a distance of several meters. Besides replacing optical bar codes, the above described characteristics of tags makes them useful in other applications as well, including access control to buildings, environmental sensing, livestock and automobile identification, inventory control, theft detection, etc. Because of their numerous applications and their low cost, RFID tags has a strong potential to become
a ubiquitous computing technology. However, several researchers have pointed out that the universal deployment of RFID systems may create new security and privacy problems (see e.g., [7, 10, 4]). The potential risks include corporate espionage, and violation of consumer (personal) privacy and location privacy. In traditional computing systems, many security and privacy problems can be solved by using cryptographic techniques [8]. Unfortunately, RFID tags are highly resource constrained and cannot support strong cryptography. To be more precise, tags could be equipped with resources to support strong cryptographic primitives, tamper resistant packaging, and other security enhancing features, but only at a higher cost of 0.5–1 USD/tag. On the other hand, significant market penetration can be expected only if tags are priced below 0.1 USD or even below 0.05 USD [7]. In this price range, tags come with the following typical characteristics:

- Storage capacity of a few hundreds bits;
- A few thousands gates available for logical functions;
- Tags are passively powered, which excludes background calculations in idle time when the tag is not "powered" by the tag reader;
- Limited distance and quality of radio transmission due to the low gain antenna and severe power constraints of the tags;
- No tamper resistance.

This means that supporting strong cryptographic primitives on low-cost tags is not a viable option today. Note that implementing even a standard cryptographic hash function, such as MD5 [5] or SHA-1 [1], is beyond the capabilities of today's tags. Hence, there is a strong need for new, lightweight cryptographic primitives that can be supported by low-cost RFID tags.

In this paper, we propose a set of extremely lightweight challenge-response type authentication protocols that can be used in low-cost RFID

systems for authenticating the tags. Tag authentication is an important primitive that serves as a fundamental building block in more sophisticated security and privacy protecting mechanisms. As a motivating example, consider a theft detection system where RFID tags are attached to stocked items, and a central tag reader periodically polls and authenticates each tag. The sys-tem should be designed in such a way that a thief cannot steal an item and go undetected by installing a cloned replacement tag that can continue authenticating itself successfully to the reader. Our protocols are designed to prevent this cloning attack.

A main contribution of our work is that we also provide an analysis of the proposed protocols. The goal of the analysis is to obtain a lower bound on the resource requirement of the best guessed attack against a given protocol, where the resource requirement is measured in the computational complexity of the attacking algorithm as well as in the number of runs that the attacker needs to observe or interfere with. Identifying a lower bound on the "cost" of the attack may help system designers to choose the best trade-off between complexity of the protocol and resistance against attacks. This is important, as the resource constraints of the tags do not allow arbitrarily complex protocols.

Outline. In Section 2, we report on some related work. In Section 3, we introduce our system model. In Section 4, we present our proposed protocols together with their analysis. Finally, in Section 5, we conclude the paper.

2 RELATED WORK

To the best of our knowledge, the first paper that calls attention to the security and privacy risks and challenges of widely deployed RFID systems is the paper of Sarma *et al.* [7]. In particular, the authors mention scarcity of tag resources as a primary challenge in providing security and privacy mechanisms in low-cost RFID systems. They suggests that new, lightweight cryptographic primitives and protocols should be developed that take into account the strong resource constraints of RFID tags. Our work is an effort in this vein. The authors of [7] also mention the problem of tag spoofing, which would enable a thief to replace an item with a cloned tag and fool a "smart shelf" that the valid item were still in stock. The tag authentication protocols we propose can be used to protect against this attack.

Several papers propose lightweight cryptographic primitives for resource constrained applications such as smart cards and sensor networks. In [2], Hoffstein *et al.* propose a lightweight public-key cryptosystem called NTRU. In [9], Stern and Stern propose a lightweight digital signature scheme. While both of those proposals lead to very efficient mechanisms compared to previously known public-key cryptosystems and digital signature schemes, they still require resources well beyond what is available on low-cost RFID tags. In [6], Perrig *et al.* propose TESLA, an efficient broadcast authentication mechanism for sensor networks, which is based on symmetric-key cryptography. However, TESLA uses hash chains and standard message authentication codes, none of which can be implemented in low-cost RFID tags. Moreover, TESLA relies on time synchronization between the base station and the sensors, which is also beyond what is feasible in low-cost RFID systems. In contrast to these proposals, our protocols use only elementary logical and arithmetical operations that can be implemented with a few gates.

In [10], Weis *et al.* propose various schemes for controlling access to RFID tags. In their proposal, each tag can be in two states: in a *locked* state, where it responds to all queries with only its meta-ID and offers no other functionality, and in an *unlocked* state, where it can perform privileged operations related to security and configuration. The goal of the proposed schemes is to ensure that the tag enters into the unlocked state only if it receives an appropriate command from a legitimate tag Hence, the proposed protocols mainly provide reader authentication. In contrast to this, in this paper, we are mainly concerned with tag authentication. In addition, the protocols proposed in [10] use standard cryptographic hash functions and one of the protocols also requires a secure pseudo-random number generator to be implemented on the tags, which seems to be infeasible with current technology.

In [3], Juels addresses the problem of privacy protection in low-cost RFID systems. He proposes a scheme where each tag stores a list of pseudonyms. Each time the tag is queried, it emits the next pseudonym from its list. In addition, the query-response rate of the tags is deliberately reduced, which guarantees that tags emit pseudonyms with a relatively low rate. Thus, an attacker can track a tag only if he has access to it for a long period of time. Due to their small storage capacity, tags can store only a short list of pseudonyms. Juels solves this problem by allowing the list to be refreshed by authorized tag readers. For this reason, mutual authentication is required between the tag and the reader (otherwise an attacker could update the list of pseudonyms in a tag). Juels proposes a lightweight mutual authentication protocol, which is based on the

release of "keys" that are supposed to be secret and associated to the parties running the protocol. After mutual authentication, the keys are refreshed; for this purpose, new keys are generated by the tag reader, and transmitted in an encrypted form to the tag. Encryption is based on a one-time pad, where the pad is selected from a series of pads maintained by the tag. The series of pads is also updated (overlayed) with new padding material in each run of the authentication protocol. The new padding material is sent in clear to the tag, but the updating procedure ensures that a pad is used (becomes "live") only after a certain number of updates. This number *m* should be chosen in such a way that an attacker cannot observe *m* consecutive updates. The rationale is that observing a run of the protocol requires physical proximity, and one may assume that an attacker cannot stay in the vicinity of a tag for an arbitrary long time.

The protocol proposed by Juels does not require the tag to perform any cryptographic operations (apart from XOR), hence it is feasible to implement it in current low-cost RFID systems. However, the protocol uses 4 messages, and updating the keys and the pads has a cost (in terms of bits that has to be pushed from the tag reader to the tag). In addition, in some applications, the assumption that the number of consecutive runs of the protocol that an attacker can observe can be upper bounded does not hold. In particular, if the tags do not move (e.g., they are attached to items stored in a depot), then the attacker can relatively easily install itself at a nearby location and eavesdrop a huge number of consecutive runs of the authentication protocol. In this case, Juels' protocol would not be appropriate for authenticating the tags. In contrast, our protocols were developed with this application in mind; on the other hand, we do not make any attempt to prevent the tracking of tags.

3 SYSTEM MODEL AND ASSUMPTIONS

We consider a system that consists of one RFID tag reader and several RFID tags. We assume that each tag shares a secret with the reader, which is established in a secure manner before the beginning of the operation of the system. Message passing between the reader and the tags is based on single-hop wireless communication. Tags are passively powered, so they can operate only if the reader provides the necessary energy. Tags are highly resource constrained: they have limited computing power, limited storage capacity, and limited communication capabilities.

We assume that the reader regularly polls the tags, each time with a new challenge, and the tags must authenticate themselves by correctly responding to the challenge. There are known protocols to do this in a secure manner, but they use standard cryptographic primitives (MAC, digital signature, encryption), which are too costly for low-cost RFID tags. We assume that computing even a standard cryptographic hash function, such as MD5 or SHA-1, exceeds the capabilities of the tags, and thus, it is infeasible in our system. In the case of standard cryptography, the speed and simplicity of an algorithm are usually qualifying factors; in our case, however, low complexity of the primitives has a first place importance.

The attacker's aim is to produce a response to a challenge. If he can do this in a feasible way, then we say that the protocol is broken. Such a success of the attacker might be achieved with or without recovering the secret key shared by the reader and the tag. This means that it is not enough to prove that the secret key cannot be calculated in a feasible way by the attacker, because he might reach his goal without it.

In the analysis of the protocols presented below, we will consider also degrading if the attacker is able to reduce the key space or can calculate some of the bits of the key in a feasible way. The data from which the attacker tries to prepare for a successful response could be obtained in a passive or in an active manner. In case of a passive attack, the attacker collects messages from one or more runs without interfering with the communication between the parties. In case of an active attack, the attacker impersonates the reader and/or the tag, and typically replays purposefully modified messages observed in previous runs of the protocol.

4 PROTOCOLS

In this section, we propose 5 lightweight tag authentication protocols that matches the stringent charac-teristics of the system described above. The presentation of each protocol is broken into three parts: first we describe the protocol and its rationale, then an analysis follows, and finally, based on the weaknesses explored by the analysis, some improvements and strengthening are suggested. Before starting the presentation of the protocols, we would like to illustrate some of the concepts on a simple example. Notations used in this example will be kept hereafter. Let us consider the following protocol:

$$R \,!\, T \;:\; x \,©\, k = a$$

$$T \,!\, R \;:\; f(x) \,©\, k = b$$

where R and T stand for the reader and the tag, respectively; k is the secret key shared by R and T ;

x is an *n* bit random challenge; and *f* is an *n*-bits to *n*-bits mapping.

The mutual information $I(h, k)$ between the observable message pair $h = (a, b)$ and the key k is $I(h,k) = H(x \oplus f(x))$, where $H(x \oplus f(x))$ is the entropy of $x \oplus f(x)$. This can be proven as follows. By definition $I(h, k) = H(h) - H(h|k)$. Because of the random selection of x, $H(h|k) = n$. Furthermore,

$$H(h) = H(a,b) = H(a, a \oplus b) = H(a \oplus b) + H(a|a \oplus b) =$$
$$= H(x \oplus f(x)) + H(x \oplus k|x \oplus f(x)) = H(x \oplus f(x)) + n$$

Thus,
$$I(h,k) = H(x \oplus f(x)) + n - n = H(x \oplus f(x)).$$

For instance, if *f* is the identity mapping (i.e., $f(x) = x$), then no information about the key can be gained from the observation of the run. However, choosing *f* as the identity mapping would trivially be a bad choice, because the attacker could simply replay the challenging message as the response. The other extreme is when *f* maps to a constant (i.e., the response message is shifted by a know constant vector). In this case, the mutual information is at maximum. Furthermore, if $x \oplus f(x)$ is a uniform mapping into an *m* dimensional subspace of the *n* dimensional vectors, then *n* - *m* bits of the key remain independently selected.

Selection of a linear binary mapping for f is dangerous. Let M and I be two n x n binary matrices, where M represents f and I is the identity matrix. The attacker can set up the following system of linear equations:

$$(M \oplus I)k = M\,a \oplus b \quad (1)$$

The solution of this system for unknown *k* is not unique if the rank of matrix *M © I* is less than *n*. Note, however, that the attacker does not have to know the exact key to be able to efficiently produce a successful response message; it is enough to know an arbitrary solution of (1).

Keeping the main structure of the protocol, possible ways of strengthening would be the following:

- *non-linearity:* use of a nonlinear *f* could make it harder for the attacker to calculate the set of pre-images (practically one way property);
- *mixed operations:* instead of the XOR operation (which is linear over binary vectors), modular integer addition or modular integer powering could be used; this would make it harder to combine the messages as well as to analyze them;
- *compression:* explicit dimension shrinking "provably" decreases the information about sensitive el-ements (key, the actual challenge) that the attacker has access to (in other words, authentication does not need to be based on invertible transformations).
- *keys:* different keys are used in the two directions.

However, strengthening must be done carefully and gradually: only light weight modifications are allowed and they must be followed by the reiteration of the analysis.

4.1 Protocol 1: XOR

Our first protocol has a structure similar to that of our example above, but it uses different keys in different directions:

$$R \to T : x \oplus k_1$$
$$T \to R : x \oplus k_2 \quad (2)$$

This would be provably secure if the keys k_1 and k_2 are selected uniformly at random in each run of the protocol. However, this clearly leads to a key establishment problem, which is further complicated by the special characteristics of our system (limited storage capacity of tags, impossibility of frequent manual key refreshing, etc.). In order to address these problems, a provably secure algorithmic key update scheme is needed, and it should also be based on a one-time pad. Note, however, that no more fresh random bits can be sent by one-time pad than that consumed for transmission!

One possibility for XOR rekeying is the following: in run *i*, *R* randomly selects a new key $k^{(i)}$ and transmits it XOR encrypted with the key *k* used in the previous run. This leads to the following protocol:

$$R \to T: a^{(i)} = x^{(i)} \oplus k^{(i)}, k^{(i)} \oplus k^{(i-1)}$$
$$T \to R: b^{(i)} = x^{(i)} \oplus k^{(0)} \quad (3)$$

where $i = 2; 3; : : :$ is a counter, which is increased by one in every new run of the protocol, $x^{(i)}$ is the *i*-th challenge, and $k^{(0)}$ and $k^{(1)}$ are initially shared secret keys. This protocol uses only XOR operations, and from this point of view it would be

ideal. However, it is breakable from two observed consecutive runs. Note that the series of keys $k^{(1)}$; $k^{(2)}$; : : : do not change randomly; the difference of them is a known value for the attacker.

Therefore, we diverge from the pure XOR protocols and we determine the consecutive session key used by R by a lightweight block stream generator with secret seeding value $k^{(0)}$, which outputs a new block in each new session. Consider the following protocol:

$$R \to T: a^{(i)} = x^{(i)} \oplus k^{(0)}$$
$$T \to R: b^{(i)} = x^{(i)} \oplus k^{(0)} \qquad (4)$$

where $k^{(i)} = \Pi(k^{(\ i_j 1)})$, and Π : $f0;\ 1g^n\ !\ f0;\ 1g^n$ is a permutation, a special key stream generator that expands seed $k^{(0)}$. Permutation Π is defined as follows.

For simpler explanation, let us assume that the key length is 128 bits (i.e., n-128). Let us cut each byte of $k^{(i-1)}$ into two half bytes. The left halves $k_{1,L}^{(i-1)}, k_{2,L}^{(i-1)}, \ldots\ldots k_{16,L}^{(i-1)}$ are concatenated into a vector denoted by $k_L^{(i-1)}$. Similarly, the right halves $k_{1,R}^{(i-1)}, k_{2,R}^{(i-1)}, \ldots\ldots k_{16,R}^{(i-1)}$ are collected into a vector denoted by $k_R^{(i-1)}$. Then, the calculation is continued in two main steps:

- Step 1: The elements of $k_R^{(i-1)}$ are permuted, where the permutation is controlled by $k_L^{(i-1)}$. The result is denoted by $k_R^{(i)}$.
- Step 2: : The elements of $k_L^{(i-1)}$ are permuted, where the permutation is controlled by $k_R^{(i-1)}$. The result is denoted by $k_L^{(i)}$.

4.2 Protocol 2: SUBSET

Consider the following protocol, where an XOR encrypted n-bit challenge is sent to the tag, which sends an m-bit portion of the challenge back as the reply:

$$R \to T: x \oplus k \qquad (8)$$
$$T \to R: f(x) = (x_{L,x_{R,[0..7]}}, x_{L,x_{R,[8..15]}}, \ldots, x_{L,x_{R,[8m..8m+7]}})$$

The challenge is divided into two parts: $x = (x_L; x_R)$. The j-th byte of x_R, denoted by $x_{R;[8j::8j+7]}$, addresses a bit of x_L, denoted by $x_{L,xR,[8j..8j+7]}$, which is considered as the j-th bit of the output vector. For concreteness and simpler presentation, let us assume the following parameters: n = 384(= 256 + 128), jx_Lj = 256, $jx_R j$ = 128 (bits), and m = 16. The probability of successful impersonation of the tag using a random response message is $2^{jm} = 2^{i16}$ for the above instance, which is considered unacceptably high in "standard" cryptography. This might not be the case for some RFID applications. Consider, for instance, our theft detection example, where the needed strength of security depends on the value of the protected goods. Taking the bit selector bytes overlapped makes it possible to increase m without increasing the length of the challenge (and the key).

One might think that, in practice, it would be easy to thwart random attacks in general, because in case of wrong responses, the reader could send an alarm signal toward the physical security subsystem. Unfortunately, it seems to be a plausible assumption that low capability devices like RFID tags are not able to run efficient forward error control mechanisms. This fact together with the poor quality of the radio channel might lead to packet error probabilities preventing the application of reliable alarms.

Passive/active attack. The attacker listens to one run of the protocol and stores messages. He guesses the first byte of k_R, therefore he also has a guess on the first byte of x_R (byte guess) and a guess on one bit of k_L (bit guess). By listening to different runs, the attacker builds a list of guesses on the bits of k_L. He continues listening to new runs until inconsistent bit guesses appear on his list (i.e., he gets both a guess of 0 and 1 for $k_{L;j}$ for some j). When the attacker gets contradicting values on the same key bit, then he can be sure that the actual guess on the first byte of k_R is wrong.

The probability distribution of the number of s-fold collisions when drawing with replacement from within a single set is known (the case of 2-fold collisions is known as the birthday paradox). For instance, the probability of two byte guesses leading to a guess of the same bit of k_L (2-fold collision) is around 0:504 if at least 19 runs are observed. Because the bit guesses are random (except when the byte guess is right), the probability that the two bit guesses obtained from incorrect byte guesses are different is 1/2. Therefore, according to the birthday paradox, the attacker detects an incorrect byte guess with probability 0:252(= 0:504=2) (from 19 observed runs). Refining the calculation by taking into account both 2- and 3-fold collisions we get, for

instance, that by observing 50 runs, the detection probability of a false byte guess improves to 0:83.

Note that the attacker can accelerate the guessing process by guessing in parallel all possible values of the first byte of k_R. He builds a matrix with 256 columns, where the columns correspond to different possible values of the key byte, and different rows correspond to different observed runs. Within a row, we have a guess for different bits of k_L. Exactly one of these key bits is guessed correctly, all others are guessed randomly. Therefore the attacker can filter out all incorrect byte guesses in parallel by finding inconsistent guesses in the corresponding rows.

Once having a reliable byte guess on the first byte of k_R, the attacker can start to explore further bits of k_L. In case of an active attack, the attacker – knowing the 1st byte of x_R – shifts the value of this byte by XOR adding the corresponding shift values to the corresponding – observed – challenge message. In this way, he can obtain all the bits of k_L from the answers by close to 256 active attacks (note that at least one bit of k_L can be obtained during the byte guess step). If an active attack is costly, then the attacker has to wait until the first byte of the randomly chosen x_R scans all (or a large portion of the) bit positions of k_L, which may lead to a number of observed runs well above 256.

The exploration of the remaining bytes of k_R needs less effort, because the attacker already knows k_L.

Strengthening. One of the main weaknesses of the above protocol is that the attacker is able to scan the bit positions of k_L through the manipulation of a single byte. Therefore, instead of only one byte, all bytes of k_R should be involved in the selection of each bit of the output. This could be done in the following way: the bits of $f(x)$ are linearly combined by bits of a separate portion of the key. In this way, the attacker is forced to guess multiple bytes of k_R.

4.3 PROTOCOL 3: SQUARING

Consider the following protocol:

$$R \rightarrow T: x$$
$$T \rightarrow R: k_L \oplus ((k_R + x)^2 \, mod 2^n)$$

where k_L and k_R are two halves of a $2n$ bit secret key $k = (k_L, k_R)$, furthermore "+" denotes integer addition.

Active attack. The attacker challenges T two times by sending $x = 0$ and $x^0 = 1$ as challenges. Then, he calculates the XOR sum of the responses:

$$z = k_R^2 \oplus (k_{R+1})^2 \, mod 2^n$$
$$= k_R^2 \oplus (k_R^2 + 2k_R + 1) \, mod 2^n$$

Note that in expression (11), two different operations are mixed.

Let b_i denote the i-th bit of b, where the LSB is b_0. Using this notation, we get that $z_0 = 1$, and $z_1 = k_{R,0}$. This means that the attacker can obtain the LSB of the key. If this key bit is zero, he can obtain also $z_2 = k_{R,1}$, and so on until he arrives at the first non-zero key bit (starting from the LSB). He cannot proceed further from this point, because he should also know the appropriate bit of k_R^2 that affects the actual carry bit.

A further helpful observation is the following: when adding two bits, there is a carry bit only in case of addition $1 \oplus 1$, which happens only for 1/4 of the combinations of bit pairs. Therefore, when we have to stop at the first non-zero bit of the key, we know that - in average - the chance that the appropriate bit of k_R^2 will not cause a carry bit is 3/4, so we proceed further in exploring more key bits under the assumption that there is no carry bit.
The above attack can be extended to the case when $x = a$ and $x' = a + 2^j$ for arbitrary a. As an example, let us consider ,

$$a = 0, j = 3 \; and \; k_R = 12_{|10} = 1100_{|2}$$

Then,

$$12^2 = 144_{|10} = 10010000_{|2} \; and \; (12 + 2^3)^2 = 400_{|10} = 110010000_{|2}$$

The XOR of the fourth bits from the right (shown in bold) gives the LSB bit of 12.

Note that if we find a few bits from k_R, we also know some bits of k_R^2, and consequently, we can learn also a few bits of k_L.

Strengthening. The use of XOR addition instead of the integer addition before squaring could be a modification to be considered. This is because (where " ¡" denotes integers subtraction), and hence, the "difference" depends on the bit which the attacker would like to find.

4.4 PROTOCOL 4: RSA

Assume an RSA encryption function E with block length n, public exponent e and secret exponent d. Consider the following protocol:

$$R \rightarrow T: x \quad (12)$$

$$T \rightarrow R: E(x \wedge k)$$

where x is a mask vector, in which there are $0 < m \cdot n$ bits set to 1 at randomly selected positions; $x \wedge k$ denotes the bitwise AND of x and k, which masks the vector k by keeping the values of the bits of k at positions where the corresponding bits in x are 1 and setting all other bits of k to 0. The number of operations when an RSA encryption function is evaluated depends on the binary weight of the exponent (according to the repeated square and multiply algorithm) as well as on the binary weight of the plain text. The second step of the protocol ensures that the binary weight of the plain text is at most m. Furthermore low weight public exponent is applied (e.g. $2^{16} + 1$).

Passive attack. The attacker can also try to determine m bits of k by listening on the channel and breaking the encryption $E(x \wedge k)$ by exhaustive search for the bits of k on coordinates determined by the mask x. This means at most 2^m trials. The workload of the attacker can be increased by increasing m, but this also increases the workload of T . Moreover, the more runs of the protocol are attacked, the more information can be obtained about k.

Active attack. The attacker changes two bit positions of an observed challenge x: a 1 bit of x is turned to 0 bit and a 0 bit of x is changed to 1 bit. In other words, a difference vector with Hamming weight 2 is XOR added to the challenge. The observation is that the response message will not change if key k had zeros at both of these two bit positions and it typically changes for the other three pairs (i.e., 01, 10, 11). Obviously, the probability of a zero pair is 1/4. In this way, the attacker has a probabilistic chance for scanning bit pairs of the key.

Strengthening. Key k is cyclically shifted with shift S before the masking operation is done, where the random shift is an appropriate mapping k' of an additional secret value and of the actual mask

$$x: S = g(k', x)$$

4.5 PROTOCOL 5: KNAPSACK

The base station maps a random challenge by a trapdoor (practically) one way function and sends the result of this mapping together with the XOR encrypted one time trapdoor value to the tag:

$$R \rightarrow T: d \oplus k, k(x, d) \quad (13)$$

$$T \rightarrow R: x \oplus k'$$

where k is an m-bit secret key and k' is an n-bit secret key, x is an n-bit challenge, and d is an m-bit trapdoor. Furthermore, $\cdot$ is a punctured multiplicative knapsack: A public set of s-bit prime numbers consisting of n primes is stored both by R and T . R selects randomly $n/2$ elements from this set and multiplies together the selected primes. The n-bit challenge x contains 1 at those bit positions which correspond to the primes selected (an order is assumed among the primes) and 0 at the remaining positions. R chooses t integers randomly from the range of 1, 2,...., s . $n/2$, and marks bits of binary representation of the product at bit positions corresponding to the selected integers. The marked bits are deleted and the binary string is shrunk in size accordingly. The resulted punctured string is the output of mapping $\cdot$. Trapdoor d consist of the integers used in puncturing, by appending these integers in order. It follows that the output of $\cdot$ has length s . $n/2$- t bits, furthermore the trapdoor is $m = t \not\in \log(s \cdot n/2)$ bits long.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed a set of extremely lightweight authentication protocols for low-cost RFID tags, and we also provided an analysis of the proposed protocols.

In general, the ultimate goal of the protocol designer is to design a protocol that has provable security under a given attacker model. Security proofs may be based on arguments from information theory (e.g., in case of one time pad), or on a reduction of the problem of breaking the protocol to a mathematical problem that is believed to be hard. Unfortunately, provable security of this kind comes with some cost: the one time pad has known key management problems, while protocols that are based on hard mathematical problems require large amount of resources even from the legal participants.

Another, more empirical design approach is to check if the protocol is resistant to the strongest known attacks against similar kind of protocols. However, the designer may have inaccurate knowledge of the capabilities of the attacker. In order to alleviate this problem, the protocol is usually oversized in dimension (e.g., number of rounds in block ciphers) and complexity. This,

however, is not desirable in low-cost RFID tags where tag resources are extremely scarce.

In this paper, we followed another approach. We built protocols from primitives that can surely be supported on low-cost RFID tags, and analyzed them, in order to see how resistant they are against various attacks. Clearly, our protocols can be broken by a powerful attacker; our goal was not to prevent this. Rather, we wanted to propose simple protocols that are amenable to analysis, and to give lower bounds on the complexity of attacking them. This allows the designer to adjust the security parameters of the system appropriately, and to find the best trade-off between security and performance. For instance, knowing a lower bound on the complexity of compromising the secret shared by the tag reader and a given tag would be useful in choosing the frequency of re-keying.

The work presented in this paper is an ongoing work. Important related problems, such as the issue of re-keying, will be addressed in future reports. We also intend to work towards a general framework in which the kind of protocols we propose here can be analyzed more systematically.

REFERENCES

[1] D. Eastlake and P. Jones. US Secure Hash Algorithm 1 (SHA1). Internet RFC 3174, September 2001.

[2] J. Hoffstein, J. Pipher, and J. Silverman. NTRU: A

[3] ring based public key cryptosystem. In ANTS III (LNCS no. 1423), pp. 267–288, 1998.

[4] A. Juels. Privacy and authentication in low-cost RFID tags. In submission. Available at http://www.rsasecurity.com/rsalabs/staff/bios/ajuels/

[5] A. Juels and R. Pappu. Squealing Euros: Privacy protection in RFID-enabled banknotes. In *Proceed-ings of the 7th Financial Cryptography Conference*, 2003.

[6] R. Rivest. The MD5 message-digest algorithm. Internet RFC 1321, April 1992.

[7] A. Perrig, R. Canetti, J. D. Tygar, and D. Song. The TESLA broadcast authentication protocol. *Cryp-toBytes*, 5(Summer), 2002.

[8] S. Sarma, S. Weis, and D. Engels. Radio-frequency identifcation: Security risks and challenges.

CryptoBytes, 6(1), 2003.

[9] B. Schneier. *Applied Cryptography*. John Wiley and Sons, 1996.

[10] J. Stern and J. Stern. Cryptanalysis of the OTM signature scheme from FC'02. In *Proceedings of the 7th Financial Cryptography Conference*, 2003.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Reliability Based Routing in Wireless Sensor Networks

E.Natarajan

Research Scholar,Dept of Computer Science, Muthayammal College of Arts and Science, Rasipuram

ABSTRACT

Sensor network deployed to recover from disaster or in military applications pose a great challenge in reliable sensed data delivery with limited power consumption. Due to the small transmission range of sensor nodes the data is forwarded using multiple hops where unexpected node failure is common at each hop. Routing techniques in sensor network gives priority to reliable transmission as the loss of important information prevents the sensor network from fulfilling its primary purpose and hence this information loss should be avoided. The commonly used routing protocols uses single path routing or multiple path routing techniques to deliver the data, without differentiating between more reliable information from less reliable and hence the overhead involved is same for all information. It leads to wastage of network resources. Thus the routing techniques evolved in sensor network should be capable of delivering the information at desired reliability at a proportionate communication cost by considering the importance of the information.

Key words: Reliability, Metrics, Wireless Sensor networks, Delivery- Success Ratio.

1.Introduction

The sensor nodes collect the data and forward it to the base station. The base station is a special node that has more computing power. Once the data is routed back to the base station, it processes the data. In a typical sensor network there are sensor nodes collecting data and base-stations with higher computing power. Data is routed back to these base-stations possibly using a "multi-hop infrastructure-less architecture".

Sensor networks are mainly deployed to monitor and provide feedback of environmental variables in areas, which are intractable to humans. With such deployments in mission critical applications, sensor networks gained importance and provide for immense potential for research in this area. Two challenging issues identified in this realm are the reduction in consumption of power by these sensors to increase their lifetime and the design of routing strategies for communication in the network.

Routing in sensor networks is a complex issue due to the large number of parameters. Unfortunately, there exists no single routing strategy, which is considered to be efficient in all aspects. A routing strategy may be shown to be efficient based on obtaining minimum load on a particular node in the grid. However, that same strategy might not be efficient if we consider load balance over all the nodes in the grid as the performance criterion. Therefore, it is better to decide the routing strategies based on the criteria of the application for which the sensor network is deployed. In addition, a routing strategy shown to be efficient in static network might not be an efficient strategy in dynamic networks. Hence, both the scenarios of static and dynamic networks are to be considered separately and also the criterion for efficiency in both scenarios is to be decided.

This paper is organized as follows. Section 2 discusses work related to existing routing protocols, path metrics and reliability improvement techniques. Section 3 presents an overview and the details of approach to the problem and the implementation of the system. Section 4 consists of the result obtained and performance evaluation of the system. Section 5 provides the conclusion with future areas of development in this area.

2. Related Work

Multi-hop routing for wireless networks [15] is and integrated routing and MAC protocol that increases the throughput of large unicast transfers in multi-hop wireless network. In this technique each node receives a packet must agree on their identities and choose one forwarder. It must have low overhead and forwarder should have lowest remaining cost to reach the ultimate destination.

Multi-path On-demand Routing protocol [8] uses hop-by-hop reliability layer for routing and data delivery in a wireless ad-hoc network. Every node makes local decision about forwarding of data packet. It also has a reliability layer incorporated that reroutes the packets through a different path if any failure occurs in delivering the packet. The reliability layer improves end-to-end packet delivery by transmitting packets limited number of times. If node fails to acknowledge packet a particular

number of times then the route that pass through that node can be deleted.

To quickly recover from failures multiple paths from source to destination are used in diffusion routing in highly resilient, energy efficient multi-path routing [10] in wireless sensor network. But the drawback is extra overhead that occurs due to multi-path formation and maintenance. Many multi-path-routing schemes are used to recover from broken paths and they are not adaptive to channel errors and required reliability specified in the packets.

3 SYSTEM ARCHITECTURE

The objective of our work is to deliver the packets at desired reliability based on the information sensed using a multi-path routing technique.

information is delivered reliably using less number of transmissions of data packet. Only limited number of paths is used between the source and the destination based on the criticality of the information to be delivered instead of using all possible paths. The three main modules of the proposed system are as follows:

- Path creation between the source and the sink
- Calculation of ETX metric
- Forwarding and retransmission of the data packets

The number of multiple paths that exists between the source and the destination varies to a great extent based on the number of nodes in the network. All these paths are used when the packets are delivered at same level of reliability. When the reliability level differs based on the information there is no need to utilize all the paths for any level of reliability. If the reliability required is small then only few paths are used and as the reliability required increases the number of paths used also increases. Such creation of multiple paths between a source and the sink has been done.

In delivering the packets in wireless sensor network routing protocols fail to find high quality paths in presence of lossy wireless links. So in routing the packets for desired reliability, a link quality metric, which considers loss in wireless links, has to be identified. The metric selected should produce high throughput in data delivery through the path. Based on the metric selected optimal path for routing has to be determined and case of any failure subsequent paths should also be selected based on the criteria identified. The metric value of all paths should be calculated.

After the creation of multiple paths between the source and the sink based on the reliability required, a reliability layer mechanism must be incorporated to deal with failures. When a node in the path through which the packets are forwarded fails, it leads to packet loss and this packet loss is overcome by using this reliability layer mechanism which routes the remaining packets through the next optimal path. It also helps in routing the lost packets to the destination.

3.1 PATH CREATION

Multiple paths between a source and a sink an be created using many techniques that are used in Ad-hoc On-demand Multi-path Distance Vector,

Multi-path Dynamic Source Routing. But these techniques create all possible paths between a source and a sink, which is not desirable when the number of paths to be used differs, based on the information.

3.1.2 Algorithm for creating multiple paths

The steps that are followed in creating multiple paths between the source and the sink is through:

- Hop Count Calculation
- Classification of Neighbor nodes
- Path Creation using control packets

3.1.2.1 Hop Count Calculation

1. Once the topology size is given it is divided into grids using the grid size specified. The number of nodes per grid is also determined. These nodes are randomly placed in the given topology.
2. Nodes that are required to act as source are attached to the sensor agent and sink node is attached to the sink agent. The simulation is started at specific time.
3. Initially sensor nodes are made to be in sleeping state and when any node has to be made working, first the node should switch to probing state and then only it switch to working state.
4. There are timers that has been set for probing, probing reply, probing reply response hello, hello reply, hello reply response and advertisement messages. When these timers expire corresponding procedures will be called.
5. Initially a node on expiration of its wakeup time switch to probe state and probe message is broadcasted to all neighbor nodes. If any node is awake within the probing range they respond with probe reply message. Based on the reply the probing node can decide whether to sleep or to work.

6. If no probe response is obtained this node starts
7. and broadcasts the hello message. The working nodes respond with hello reply message, which contains the information stored in its sink table.
8. Sink broadcasts the advertisement message with the value of hop count has 1 to all neighbor nodes. On receiving this advertisement message, each node increments the hop count value that is specified in the message and store it in its sink .

3.1.2.2 Classification of Neighbor Nodes

1. The maximum number of neighbors for each node is limited to 1000.
2. When a node receives an advertisement packet, it stores the source of the packet as its neighbor. Thus each node maintains a list of its neighboring nodes.
3. A node classifies its neighbors by comparing its hop count value with that of its neighbors as follows:
   - 3.1.1. If the hop count value of the neighbor node is less by one, then this neighbor node is placed in a negative set.
   - 3.1.2. If the hop count value of the neighbor node is greater by one, then this neighbor node is placed in a positive set.
   - 3.1.3. If the hop count value of the neighbor node is the same, then this neighbor node is placed in a zero set.
4. The classified nodes are added to the appropriate fields (positive, negative, zero) in the sink table.
5. A Count of number of nodes in each set is also maintained in the sink table.

3.1.2.3 Path Creation using Control Packets

On generating a packet, the source node determines the importance of the information it contains and decides the desired reliability r_s for it. It also knows the local channel error e_s. The source s is at distance hs from the sink, which is computed in hop count calculation step. The neighborhood set of the source is divided into 3 subsets H_s^- , H_s^0 , H_s^+, which contain neighbors at distance hs −1,hs, hs + 1 respectively with help of previous step. Using these values, the source computes the number of paths N_s, required for delivering the packet at desired reliability to the sink.

3.2 CALCULATION OF ETX METRIC

The metric most commonly used by existing ad hoc routing protocols is minimum hop-count. These protocols typically use only links that work well enough to deliver routing probe packets (query packets, as in DSR or AODV, or routing updates, as in DSDV). This approach implicitly assumes that links either work well or don't work at all.

Minimizing the hop-count maximizes the distance traveled by each hop, which is likely to minimize signal strength and maximize the loss ratio. Even if the best route is a minimum hop-count route, in a dense network there may be many routes of the same minimum length, with widely varying qualities; the arbitrary choice made by most minimum hop-count metrics is not likely to select the best.

The solution that overcomes the drawbacks of minimum hop- count metric is ETX metric. ETX finds paths with the fewest expected number of transmissions (including retransmissions) required to deliver a packet all the way to its destination. The metric predicts the number of retransmissions required using per-link measurements of packet loss ratios in both directions of each wireless link. The primary goal of the ETX design is to find paths with high throughput, despite losses. The metric's overall goal is to choose routes with high end-to-end throughput. The metric must account for the following issues:

- The wide range of link loss ratios.
- The existence of links with asymmetric loss ratios.
- The interference between successive hops of multi-hop paths.

3.2.1 ETX Metric

The ETX of a link is the predicted number of data transmissions required to send a packet over that link, including retransmissions. The ETX of a route is the sum of the ETX for each link in the route.

ETX has several important characteristics:

- ETX is based on delivery ratios, which directly affect throughput.
- ETX detects and appropriately handles asymmetry by incorporating loss ratios in each direction.
- ETX can use precise link loss ratio measurements to make fine-grained decisions between routes.

- ETX penalizes routes with more hops, which have lower throughput due to interference between different hops of the same path.

- ETX tends to minimize spectrum use, which should maximize overall system capacity.

3.2.2 Steps followed in calculating ETX metric

1. A timer is scheduled to expire at an interval of 0.1 second and each node is assigned a counter with an initial value of 0.
2. On expiration of this timer probe packets are transmitted to neighboring nodes and the timer is rescheduled.
3. When a node receives the probe packet its counter value is incremented by one.
4. The above two steps are repeated for w seconds.
5. At time t, node calculates the delivery ratio by knowing the number of the probes it received during the last w seconds and the actual number of the probes that should have been received.
6. ETX metric of the link is calculated using the delivery ratio obtained in previous step.
7. The steps 1 to 6 are repeated for all links in a path and the summation of each link in the path gives the path ETX metric value.
8. When ETX values of all paths are determined, the path that has least ETX metric value is selected as an optimal path.

Each node's ETX value is the sum of the link ETX values along the lowest-ETX path to E. A link's ETX value is the inverse of the link's delivery probability in the forward direction.

3.3 FORWARDING AND RETRANSMISSION

Reliability layer mechanism that improves reliability in sensor network can be done making use of techniques like:

- Link-layer retransmission
- Blacklisting bad links
- End-to-end routing metrics

The main algorithm used is as follows:

The data packets are forwarded through the selected optimal path using data forwarding algorithm described below.

1. On path failure the remaining and the lost packets are sent through the next available optimal path using the retransmission algorithm.
2. At the end of the simulation, the success ratio, which is defined as the number of packets received successfully at the sink to the number of copies of packets actually used is calculated.

Algorithm for forwarding the data packet:

1. The data timer present in the source is scheduled for an interval. When the time interval expires a data packet is sent to the next node prevailing in the path.
2. The timer is rescheduled whenever a data packet is sent from the source.
3. The node on receiving the data packet passes it to its neighbor node which is present in the path.
4. Whenever a data packet is received by the sink, it sends an acknowledgement to the source.
5. This process continues until all the packets are delivered to the sink or the source detects the failure in the path and stops the transmission.

Algorithm used for retransmission:

1. Path failure is detected whenever acknowledgement is not received by the source for considerable amount of data packets.
2. The sent packets for which the acknowledgements are not received are categorized as lost packets and the packets, which are yet to be sent, are categorized as remaining packets.
3. When a path failure is detected the existence of optimal paths are found. If exists, the remaining data packets are sent to the sink through the next best optimal path. This process too makes use of the data timer, which is used in the forwarding.
4. A retransmission timer is scheduled for sending the lost packets. This works similar to that of the data timer.
5. Thus the remaining packets and the lost packets are sent to the sink based on the reliability level successfully.

4. PERFORMANCE EVALUATION

The performance of our system is evaluated mainly based on the following parameters.

- Delivery Ratio
- Success Ratio

4.1 Delivery Ratio

The Delivery Ratio (DR) service level parameter reports the networks effectiveness in transporting an offered packet load in one direction of a single connection. The DR is a ratio of successful packet receptions to attempted packet transmissions. Attempted packet transmissions are referred to as Sent Packets. Successfully delivered packets are referred to as Received Packets. These loads may be further differentiated as being within the committed information rate or as burst excess.

Method:

1. For each simulation run, sum up the total number of packets sent and the total number of packets received for that run.
2. Calculate the packet delivery ratio for each run as:

$$DeliveryRatio = \frac{TotalPacekts\ \text{Re}\ ceived}{TotalPacketsSent}$$

4.2 Success Ratio

Success Ratio (SR) is intended to provide a metric, which summarizes the overall service provided. This metric captures the cost to send data, and includes retransmissions. This metric is computed by normalizing the total number of data transmissions by the number of messages sent to reflect the cost of packets that are sent.

$$SuccessRatio = \frac{TotalPackets\ \text{Re}\ ceived}{TotalNumberofCopiesofPacketUsed}$$

4.3 RESULTS

A detailed simulation of the modified multi-path routing protocol - has been done varying the reliability, as discussed in the previous sections.

Number of Nodes : 300
Hop Distance between Source and Sink : 6
Error Rate : 20%
Source : Node 12
Sink : Node 300
Reliability : 70%
Number of Paths to be created : 4

1. Paths Created

Path 1: 12->27->93->194->291->287->300
Path 2: 12->80->172->177->255->277->300
Path 3: 12->31->96->194->180->270->300
Path 4: 12->79->164->165->196->285->300

2. ETX Values

Path 1: 12->27->93->194->291->287->300 has ETX value 2.381543
Path 2: 12->80->172->177->255->277->300 has ETX value 2.251242
Path 3: 12->31->96->194->180->270->300 has ETX value 3.204246
Path 4: 12->79->164->165->196->285->300 has ETX value 2.855913

2.1.1 Optimized Paths

1. Path 2 with ETX Value 2.251242
2. Path 1 with ETX Value 2.381543
3. Path 4 with ETX Value 2.855913
4. Path 3 with ETX Value 3.204246 First best Path is Path 2

4.3.1 Delivery Ratio and Success Ratio Results

Also detailed comparisons between the two protocols are done for performance evaluation and are given as plots. Figure 4.1 and Figure 4.2 show the performance comparison of modified multi-path routing protocol and Gradient Broadcast.

The Reliability Vs. Number of Paths with maximum reliability of 90%. It is evident from the graph that as the reliability level increases the number of paths required also gets increased.

5. References

[1] Adya A., P. Bahl, J. Padhye, A. Wolman, and L. Zhou (2004), 'A multi-radio unification protocol for IEEE 802.11 wireless networks', In BroadNets.

[2] Akyildiz I.F., Su W., Sankarasubramaniam Y.

and Cayirci E. (2002), 'Wireless sensor networks: a survey', Computer Networks, 38(4): 393-422.

[3] Bhatnagar S, Deb B., and Nath B. (2001), 'Service differentiation in sensor networks', In Wireless Personal Multimedia Communications.

[4] Budhaditya Deb, Sudeept Bhatnagar and Badri Nath (2003),Reliable Information Forwarding Using Multiple Paths in Sensor Networks', In Proceedings of the 28th Annual IEEE International Conference on Local Computer Networks (LCN'03).

[5] Clark B. N., Colbourn C. J., and Johnson D. S. (1990), 'Unit Disk Graphs', Discrete Mathematics, 86(1-3): 165-177.

[6] David B Johnson and David A Maltz (1996), 'Dynamic source routing in ad hoc wireless networks', In Imielinski and Korth, editors, Mobile Computing, volume 353. Kluwer Academic Publishers.

[7] De Couto, Daniel Aguayo (2003), 'A High-Throughput Path Metric for Multi-Hop Wireless Routing', In MOBICOM, pages 14–19.

[8] Edoardo Biagioni, Shu Hui Chen (2004), 'A Reliability Layer for Ad-Hoc Wireless Sensor Network Routing', In Proceedings of the 37th Hawaii International Conference on System Sciences.

[9] Fan Ye, Gary Zhong, Songwu Lu and Lixia Zhang (2005), 'GRAdient Broadcast: A Robust Data Delivery Protocol for Large Scale Sensor Networks', *ACM* Wireless Networks (WINET), 11(2).

[10] Ganesan D., Govindan R., Shenker S., and Estrin D. (2002), 'Highly resilient, energy-efficient multi-path routing in wireless sensor networks', In Mobile Computing and Communications Review (MC2R) Vol 1., No. 2.

[11] Intanagonwiwat C., Govindan R., and Estrin D. (2000), 'Directed diffusion: a scalable and robust communication paradigm for sensor networks', In MOBICOM, pages 56–67.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Deep Packet Inspection (DPI) for Internet and Surveillance

[1]Mrs.S.Thilagavathi M.C.A, M.Phil., [2]Ms.N.Nithya M.Sc.,

[1]Head & Assistant Professor in Comp Science [2]Assistant Professor in Comp.Science
[1 & 2] Terf's Academy College of Arts & Science, Tirupur, Tamilnadu, India.

ABSTRACT

Deep packet inspection (DPI) analyses all data of a packet (headers and payload) as it passes an inspection point in order to determine the application protocol and/or application transported. In this paper we present how the DPI works and the challenges faced in implementation. Even though various sorts of sectors disagree with DPI still it is effectively implemented in various confidential activities. We make an effective comparison of DPI with other existing traditional technologies.

Keywords: Deep packet inspection, Payload, Applications, Technologies, Data traffic

1. Introduction

Deep Packet Inspection (DPI) (also called complete packet inspection and Information eXtraction - IX -) is a form of computer network packet filtering that examines the data part (and possibly also the header) of a packet as it passes an inspection point, searching for protocol non-compliance, viruses, spam, intrusions, or defined criteria to decide whether the packet may pass or if it needs to be routed to a different destination, or, for the purpose of collecting statistical information. [2]

There are multiple ways to acquire packets for deep packet inspection. Using port mirroring (sometimes called Span Port) is a very common way, as well as optical splitter. Deep Packet Inspection (and filtering) enables advanced network management, user service, and security functions as well as internet data mining, eavesdropping, and internet censorship. Although DPI technology has been used for Internet management for many years, some advocates of net neutrality fear that the technology may be used anticompetitively or to reduce the openness of the Internet.[3] DPI is currently being used by the enterprise service providers, and governments in a wide range of applications.[4]

Deep packet inspection is a networking technology that businesses and Internet Service Providers use to monitor what applications are generating and receiving network traffic. Data flows across the Internet as packets, and these packets are composed of two key elements: headers and payloads. The header directs packets to their terminal destinations, like an address on a postcard directs to postcard to the recipient's address. The payload holds the packet's actual contents; using the postcard metaphor, the payload holds the image, text, color of the text, handwriting style, and so forth. Whereas earlier networking technologies analyzed and filtered packets based on header information, DPI systems permit a more granular analysis of packets based on their payload. Further, DPI can modify the contents of packets and identify data traffic even when it is encrypted. The rest of this paper is organized as follows: section 2 and 3 introduces an overview of goals are using the deep packet inspection. Section 3 introduces the comparison among existing technologies, section 4 and 5 presents the search algorithms and applications respectively. Finally we conclude with section 6.2. Deep Packet Inspection. All traffic on the Internet travels around in what is called an IP packet. An IP packet is a string of characters moving from computer A to computer B. On the outside of this packet is the address where it is being sent. On the inside of the packet is the data that is being transmitted.

The string of characters on the inside of the packet can be conceptually thought of as the "payload," much like the freight inside of a railroad car. These two elements, the address and the payload, comprise the complete IP packet.

When you send an e-mail across the Internet, all your text is bundled into packets and sent on to its destination. A deep packet inspection device literally has the ability to look inside those packets and read your e-mail (or whatever the content might be).

Products sold that use DPI are essentially specialized snooping devices that examine the content (pay load inside) of Internet packets. Other terms sometimes used to describe techniques that examine Internet data are packet shapers, layer-7 traffic shaping, etc.[5]

3. Goals of DPI

The design and implementation of the deep packet inspection has several challenges which harden its advancement process. In this section, we list the different Process.

3.1 The search algorithm complexity: the complexity of the algorithm and the operations of comparison against the signatures of intruder de-crease the throughput of the system. Thus, search algorithms are the main focus point in DPI researches, whereas matching process is resource consuming

3.2 Increasing number of intruder signature: according to the verity of attacks, the needs for new intruder signature increase. Therefore, the large number of signatures makes the task of IDS harder whereas the matching process must inspect tracing against all attacks.

3.3 The overlapping of signatures: the signatures of attacks usually are not general so the

Signatures can be categorized into groups according to common properties like protocol type.

3.4 The Location of signature unknown: due to variety types of attacks on different types of applications, the pattern of intruders is not localized in specific place in the packet which means that the IDS must inspect all the payload of the packet against the attacker signatures.

3.5 Encrypted Data: the data which is encrypted cannot be inspected by DPI. However, there are some solutions to overcome this problem by plugging the DPI component behind the decryption device. The DPI system as we mentioned before has many challenges and in the same time it have to provide the requirements for network need.

4.Comparison with existing technologies

Three technologies used widely for surveillance are Shallow packet inspection, Medium packet inspection and Deep packet inspection. In this Shallow packet inspection examines the packets header file information. This inspects the layer 1-3 in OSI model and it can block data that comes from certain IP address (e.g: spam or viruses). Medium packet inspection analyses data in layers 1-6. This includes the presentation layer of OSI, in which file formats are defined.MPI technology can block certain type of files in network management. Deep packet inspection surveillance method can identify the origin and content of each packet of data. They can monitor data at the OSI layers 1-7. DPI surveillance technologies can also able to monitor the traffic of network data that is sent over the Internet at all seven layers.

5. Algorithms

The matching for pattern depends on the algorithmic way to process the data and return the result of existence of the pattern or not in considerable time. The first factor is the computation operations to make comparison between the pattern and the data and second is the number of patterns that need to be compared with the traffic of the incoming data. Historically, the first string matching algorithm was the brute force (BF) algorithm which compares the first character in the pattern with the data stream. If a single character match, BF compares it with the next character of the pattern and so on. Finally, if the whole pattern is finished, it issues the pattern matching results. Later on, many algorithms appear to increase the performance of matching. These algorithms can be categorized according to the implementation as soft-ware based, HW based or mixture of both implementations. Briefly, there are a lot of algorithms for pattern matching. However, the most famous software based algorithms are Knuth-Morris-Pratt (KMP) .We will summarize the concept behind selected algorithms and their implementation for DPI.

KMP Algorithm: The Knuth-Morris-Pratt(KMP) algorithm came as an enhancement for the brute force algorithm which was we introduced before as the early work for pattern matching. The improvement of KMP over the BF is performed by skipping characters when the mismatch occurs in the comparison phase. This skipping for characters depends on preprocessing phase of KMP to the patterns. The result of the KMP is somehow similar to the finite automata for patterns representation in which depending on every match and mismatch a certain jump over the input stream occurs. Additionally, KMP and BM algorithms are designed for single pattern searching.If the pattern length is m bytes, the complexity of the matching algorithm will be of O(m + n) matching this pattern in an n bytes stream. If there are k patterns, the search time will be O(k(m + n)) according to that the single search is performed k times.

6. Applications

Deep packet inspections have been extensively used by the European, china and other nations. Government of china used DPI and has blocked various sites in china.The Chinese government uses Deep Packet Inspection to monitor and censor network traffic and content that it claims is harmful to Chinese citizens or state interests. This material includes pornography, information on religion, and

political dissent. Chinese network ISPs use DPI to see if there is any sensitive keyword going through their network. If so, the connection will be cut. Other than china countries like USA, Iran and Malaysia adopted this DPI for their surveillance in politics and security. DPI is appreciated for its features such as net neutrality and its surveillance of file sharers.[6][7]

7. Conclusion

Today's Internet is guided by network neutrality, a principle that suggests that routers read packet headers and subsequently shuttle packets around the Internet on a best-effort basis. This idea is juxtaposed against a system where content is examined and data packets (de)prioritized based on packet contents. While DPI may have some benefits for network diagnostics and security, it is more substantively used to interfere with and modify data transmissions for advertising purposes, as well as to establish differential data traffic queues. Deep packet inspection is on the active area of research. Hence, an effective technique can be developed in future to enhance security and reliability with an improved DPI surveillance method.

8.References

[1] Tamer AbuHmed, Abedelaziz Mohaisen, and DaeHun Nyang, "Deep Packet Inspection for Intrusion Detection Systems: A Survey", Information Security Research Laboratory, Inha University, Incheon 402-751, Korea.

[2] Dr. Thomas Porter (2005-01-11). "The Perils of Deep Packet Inspection". Security Focus. Retrieved 2008-03-02.

[3] Hal Abelson, Ken Ledeen, Chris Lewis (2009). "Just Deliver the Packets, in: "Essays on Deep Packet Inspection", Ottawa". Office of the Privacy Commissioner of Canada. Retrieved 2010-01-08.

[4] Ralf Bendrath (2009-03-16). "Global technology trends and national regulation: Explaining Variation in the Governance of Deep Packet Inspection, Paper presented at the International Studies Annual Convention, New York City, 15–18 February 2009".

[5]Http://netequalizernews.com/2011/02/08/what-is-deep-packet-inspection-and-why-the-controversy/

[6] Genny Pershing. "Network Neutrality: Insufficient Harm". Cybertelecom. Retrieved 2008-06-26.

[7] Sailesh Kumar, "Content Aware Networks", Cisco Research.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A Survey on Steganography Using DCT & DWT Techniques

[1]K.Swathi [2]G.Sathyavathy

[1]Research Scholor [2]Assistant Professor
Sri Ramakrishna College of Arts and Science for Women,Coimbatore-44

ABSTRACT

Steganography is the art of hiding the fact that communication is taking place, by hiding information in other information. Many different carrier file formats can be used, but digital images are the most popular because of their frequency on the Internet. Steganography is a set of techniques for hiding or encoding message or data within an image. In image Steganography, secret communication is achieved to embed a message into cover image and generate a stego-image.

Index Terms - Steganography, data hiding, image steganography, transform domain technique

I. INTRODUCTION

Steganography is the art of hiding information through original files that the existence of the message is unknown. The term steganography is comes from Greek word Steganos, which means, “Covered Writing”. The original files can be referred to as cover text, cover image, or cover audio. After inserting the secret message it is referred to as stego-medium. A stego-key is used for hiding process to control detection and recovery of the embedded data. There are two techniques available to achieve this goal. One is cryptography, where the sender uses an encryption key to encrypt the message. This encrypted message is transmitted through the insecure public channel, and decryption algorithm is used to decrypt the message. The restoration of the original message is achieved only if the receiver has the decryption key. The second method is steganography, where the secret message is inserted in another medium. While cryptography protects the content of messages, steganography hides the message so that intermediate persons cannot see the message.

The purpose of cryptography is to secure communications by changing the data into a form that cannot be understood. Steganography techniques, hide the existence of the message itself, which makes it difficult for a third person to find out where the message is. Sometimes sending encrypted information may draw attention, while invisible information will not. Accordingly, cryptography is not the good result for secure communication; it is only part of the solution. Both techniques can be used together for better protection of message. In this case, even if steganography fails, the message cannot be recovered because a cryptography technique is used as well. The cracking of steganographic messages is called steganalysis.

The paper is organized as follows: Section 2 provides the motivation for developing a routing attack, while Section 3 describes the architecture and implementation of proposed model. The performance

of survey is evaluated in Section 4 and the paper concludes in Section 5.

II. Steganography Techniques

There are several Steganographic techniques for image file format which are as follows:

- Spatial domain technique
- Masking and filtering
- Transform techniques
- Distortion Techniques

A. Spatial Domain Technique:

There are many versions of spatial steganography, all directly change some bits in the image pixel values in hiding data. Least significant bit (LSB)-based steganography is one of the simplest techniques that hides a secret message in the LSBs of pixel values without introducing many detectable distortions. To our human eye, changes in the value of the LSB are invisible. Embedding of message bits can be done either sequentially or randomly. Least Significant Bit (LSB) replacement, LSB matching, Matrix embedding and Pixel value are some of the spatial domain techniques. There is less chance for degradation of the original image. More information can be stored in an image.

B. Masking and Filtering:

These techniques hide information by marking an image. These embed the information in the more significant areas than just hiding it into the noise level. The hidden message is more important to the cover image. Watermarking techniques can be applied without the fear of image destruction due to lossy compression as they are more incorporated into the image. This method is much more robust than LSB replacement with respect to compression since the information is hidden in the visible parts of the image.

C. Transform Domain Technique:

This is a more complex way of hiding information in an image. The process of embedding data in the frequency domain of a signal is much stronger than embedding principles that operate in the time domain. Most of the strong steganographic systems today operate within the transform domain Transform domain techniques have an advantage over LSB techniques as they hide information in areas of the image that are less exposed to compression, cropping, and image processing. Transform domain techniques are broadly classified into :

- Discrete Fourier transformation technique (DFT).
- Discrete cosine transformation technique (DCT).
- Discrete Wavelet transformation technique (DWT).

D. Distortion Techniques:

Distortion techniques require awareness of the original cover image during the decoding process where the decoder functions to ensure for differences between the original cover image and the distorted cover image in order to restore the secret message. The message is encoded at pseudo-randomly chosen pixels. In any steganographic technique, the cover image should never be used more than once. If an attacker interferes with the stego-image by cropping, scaling or rotating, the receiver can easily detect it. In some cases, if the message is encoded with error correcting information, the change can even be reversed and the original message can be recovered.

III. RELATED WORKS

Rajkumar Yadav et al. [6] proposed a new image steganography technique for embedding messages into gray level Images. This new technique distributes the message uniformly throughout the cover image. The image is divided into blocks of equal sizes and the message is then inserted into the central pixel of the block using cyclic combination of 6th, 7th & 8th bit. The blocks of the cover image are chosen randomly using the pseudo random generator seeded with a secret key. Cyclic combination of last three bits of pixel value provides 100% chances of message insertion at the pixel value and division of image into blocks distribute the message uniformly throughout the image.

Dr. Ekta Walia et al., [2] provides analysis of Least Significant Bit (LSB) based steganography and Discrete Cosine Transform (DCT) based steganography. LSB based steganography insets the text message in lsb of digital data. Converting an image from a format like BMP or GIF which reconstructs the original message exactly to a JPEG which does not and then back could destroy the information hidden in the LSBs. DCT based steganography embed the text message in lsb bits of the discrete cosine (DC) coefficient of digital picture.

Jonnson et al., [9] proposed a new method for the most LSB for hiding data in digital image. Another technique for hiding data in digital image is GLM technique. Main drawback of above methods is that if the intruder changes least significant bit (LSB) of all image pixels then hidden message can be destroyed. In this case the change in image quality is in the range of +1 to -1 at each pixel position that. Second drawback is that LSB bit may be corrupted by hardware imperfections or quantization noise due to which message can be distorted. In GLM method if lsb bit changes due to above two problems then pixel value become from even to odd or odd to even due to which message can be destroyed.

Zhang et al., [3] proposed a method for attacking PVD steganography based on observing the histogram of the prediction errors. Since \0" and \1" are equally distributed in the binary secret data, the occurrence of the decimal values are also equally distributed in each Rk. The reason is very similar to the LSB steganography.

Katzenbeisser et al., [10] describe a method of substitution that attempts to be more robust than simple bit substitution by reducing the statistical variations in the stego-image in an attempt to avoid detection. The technique divides the cover image into several disjoint cover regions. To embed the secret information, *l(m)* disjoint cover regions are selected to encode one-bit each in the parity of the cover image. If the parity does not match the secret bit to be encoded, then one of the LSB values is flipped.

Yang et al*.,* [5] proposed an adaptive LSB substitution based data hiding method for image. To achieve better visual quality of stego-image it takes care of noise sensitive area for embedding. Proposed method differentiates and takes advantage of normal texture and edges area for embedding. This method analyzes the edges, brightness and texture masking of the cover image to calculate the number of k-bit LSB for secret data embedding. The value of k is high at non-sensitive image region and over sensitive image area k value remain small to balance overall visual quality of image. The LSB's (k) for embedding is

computed by the high-order bits of the image. It also utilizes the pixel adjustment method for better stego-image visual quality through LSB substitution method.

Madhu et al., [6] proposed an image steganography method, based on LSB substitution and selection of random pixel of required image area. This method is target to improve the security where password is added by LSB of pixels. It generates the random numbers and selects the region of interest where secret message has to be hidden. The strength of method is its security of hidden message in stego-image, but has not considers any type of perceptual transparency.

E. Discrete Wavelet transform

The Wavelet Transform provides a time-frequency representation of the signal. The Discrete Wavelet Transform (DWT) of image signals produces a non-redundant image representation, which provides better spatial and spectral localization of image formation, compared with other multi scale representations such as Gaussian and Laplacian pyramid. Recently, Discrete Wavelet Transform has involved more and more interest in image de-noising. The DWT can be interpreted as signal decomposition in a set of independent, spatially oriented frequency channels. The signal is passed through two complementary filters and emerges as two signals, approximation and Details. This is called decomposition or analysis. The components can be assembled back into the original signal without loss of information. This process is called reconstruction or synthesis. The mathematical manipulation, which implies analysis and synthesis, is called discrete wavelet transform and inverse discrete wavelet transform. An image can be decomposed into a sequence of different spatial resolution images using DWT.



Fig:1 Original Image



Fig:2 Decompressed Image

F. Discrete cosine transformation technique

DCT is an accurate and robust face recognition system and using certain normalization techniques, its robustness to variations in facial geometry and illumination can be increased. An alternative holistic approach to face recognition is discrete cosine transform. Namely, an affine transformation was used to correct scale, position, and orientation changes in faces. It was seen that tremendous improvements in recognition rates could be achieved with such normalization. Illumination normalization was also investigated extensively. Various approaches to the problem of compensating for illumination variations among faces were designed and tested, and it was

concluded that the recognition rate of the specific system was sensitive to many of these approaches. This was because the faces in the databases used for the tests were uniformly illuminated and these databases contained a wide variety of skin tones. That is, certain illumination normalization techniques had a tendency to make all faces have the same overall grey-scale intensity, and they thus resulted in the loss of much of the information about the individuals' skin tones.



Fig:3 Original Image



Fig:4 Decompressed Image

IV. CONCLUSION

This paper gave an overview of different steganographic techniques and also critically analyzed different proposed techniques which show that visual quality of the image is degraded when hidden data increased up to certain limit using LSB based methods. Transform domain technique were analyzed with different images.

REFERENCES

[1] Katzenbeisser, S., Petitcolas, F., "Information Hiding Techniques for Steganography and Digital Watermarking", Artech House, Inc., Norwood, MA, 2000.

[2] Dr. Ekta Walia a, Payal Jainb "An Analysis of LSB & DCT based Steganography ",Global Journal of Computer Science and Technology, 4 Vol. 10 Issue 1 (Ver 1.0), April 2010, p4-8

[3] X. P. Zhang and S. Z. Wang, Vulnerability of pixel-value diRerencing steganography to histogram analysis and modication for enhanced security, Pattern Recognition Letters, vol. 25, no. 3, pp. 331-339, 2004.

[4] STEGANOGRAPHY WITH UNIFORM DISTRIBUTION OF MESSAGE ", Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November 2011, p29-43

[5] H. Yang, X. Sun and G. Sun, "A High-Capacity Image Data Hiding Scheme Using Adaptive LSB Substitution", Journal: Radioengineering, vol. 18, no. 4, (2009), pp. 509-516.

[6] V. Madhu Viswanatham and J. Manikonda, "A Novel Technique for Embedding Data in Spatial Domain", International Journal on Computer Science and Engineering, IJCSE, vol. 2, (2010).

[7] Rajkumar Yadav, Ravi Saini and Kamaldeep "A New Image Steganography Approach for Information Security Using Gray Level Images in Spatial Domain", Vol. 3 No. 7 July 2011, ISSN: 0975-3397, p2679-2690.

[8] M. Chaumont and W. Puech, "DCT-Based Data Hiding Method To Embed the Color Information in a JPEG Grey Level Image", 14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, copyright by EURASIP, (2006) September 4-8.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

RECOGNIZING PRE AND POST SURGERY FACES USING MULTIOBJECTIVE EVOLUTIONARY FIREFLY ALGORITHM

A.Nirmala [1] K.Dilipkumar [2]

[1] Asst. Professor, Department of Computer Applications,

Dr. N.G.P. Arts and science college, Coimbatore , India

[2]M.Phil Scholar , Dr. N.G.P. Arts and science college, Coimbatore, India,

ABSTRACT

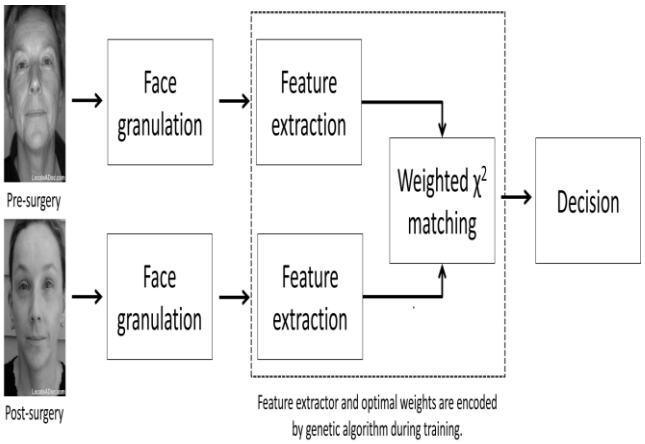
Plastic surgery procedures provide a proficient and enduring way to enhance the facial appearance by correcting feature anomalies and treating facial skin to get a younger look. Apart from cosmetic reasons, plastic surgery procedures are beneficial for patients suffering from several kinds of disorders caused due to excessive structural growth of facial features or skin tissues. The security and privacy problems has been overcome. There is a possibility of attacking others privacy by Criminals or attackers using plastic surgery technology.Our proposed method deals better for overcoming such problems. Transmuting facial geometry and texture increases the intraclass variability between the pre- and post-surgery images of the same individual. Therefore, matching post-surgery images with pre-surgery images becomes an arduous task for automatic face recognition algorithms. Variations in pose, expression, illumination, aging and disguise are considered as major challenges in face recognition and several techniques have been proposed to address these challenges. In Existing, multi-objective evolutionary granular algorithm is used to match face images before and after plastic surgery. The problem faced by using existing system is it fails to maintain diversity in a population which in terms decreases quality of solution. To overcome from these problems our proposed method uses Firefly algorithm for premature and population diversity problem. On the plastic surgery face database, the proposed algorithm yields high identification accuracy as compared to existing algorithms and a commercial face recognition system.

I. INTRODUCTION

Face recognition system

A **facial recognition system** is a computer application for automatically identifying or verifying a person from a digital image or a video frame from a video source. One of the ways to do this is by comparing selected facial features from the image and a facial database. Facial recognition has received significant attention in the last few years and is increasingly being used for both identification (1: n) and verification (1:1) on large identity projects across the public and private sector. Facial recognition analyses characteristics of a person's face image input through a camera and can be broadly classified into static and dynamic/video matching. Facial recognition systems at a very high level work by recognising a human face from scene and extract it. The system then measures nodal points on the face, distance between eyes, shape of the cheekbones and other distinguishable features. These nodal points are then compared to the nodal points computed from a database of pictures in order to find a match.

Plastic surgery procedures provide a proficient and enduring way to enhance the facial appearance by correcting feature anomalies and treating facial skin to get a younger look. Apart from cosmetic reasons, plastic surgery procedures are beneficial for patients suffering from several kinds of disorders caused due to excessive structural growth of facial features or skin tissues. Plastic surgery



procedures amend the facial features and skin texture thereby providing a makeover in the appearance of face.

Facial aging is a biological process that leads to gradual changes in the geometry and texture of a face. Unlike aging, plastic surgery is a spontaneous process that is generally performed contrary to the effect of facial aging. Since the variations caused due to plastic surgery procedures are spontaneous, it is difficult for face recognition algorithms to model such non-uniform face transformations. On the other hand, disguise is the process of concealing one's identity by using makeup and other accessories. Both plastic surgery and disguise can be misused by individuals trying to conceal their identity and evade recognition. Variations caused due to disguise are temporary and reversible; however, variations caused due to plastic surgery are long-lasting and may not be reversible. Owing to these reasons, plastic surgery is now established as a new and challenging covariate of face recognition alongside aging and disguise.

Research presents a multi-objective evolutionary granular computing based algorithm for recognizing faces altered due to plastic surgery procedures. Proposed algorithm starts with generating non-disjoint face granules where each granule represents different information at varying size and resolution. Further, two feature extractors, namely Extended Uniform Circular Local Binary Pattern (EUCLBP) and Scale Invariant Feature Transform (SIFT) are used for extracting discriminating information from face granules. Finally, different responses are unified in an evolutionary manner using a multi-objective genetic approach for improved performance. The performance of the proposed algorithm is compared with a commercial-off-the-shelf face recognition system (COTS) for matching surgically altered face images against large scale gallery.

Figure 1: Example of the face in research

Problem Definition

Multi-objective evolutionary granular algorithm is used to match face images before and after plastic surgery. The problem faced by using this approach is it fails to maintain diversity in a population which in terms decreases quality of solution.

Objective

Different aspects related to plastic surgery and face recognition has to be investigated. The mean objective of this research is to recognize faces before and after surgery using multi-objective granular algorithm the face feature to be extracted. The extracted feature can be selected and optimized. Finally Learning strategies can be made for those selected and optimized features.

Scope

In this research the scope lies in recognize pre surgery face for the corresponding post surgery face. The face feature can be extracted .The feature selection and optimization task has to be done by Multi-objective Firefly Algorithm. This selects best features and calculates weight value. The optimization process is considered as best when compare with the previously used Optimization methods.

Organisation

This research work can be organised as follows. Literature work of this research has been discussed in the Chapter2.The Methods used for feature extraction, Feature selection and optimisation has been discussed in the Chapter 3.System Implementation of this research has been discussed in Chapter 4.Experimental results and Discussion has been done in Chapter 4.Summarization of the research has been described in Chapter 5.

2.PROPOSED METHODOLOGY

Face image granulation

Let F be the detected frontal face image of size n × m. Face granules are generated pertaining to three levels of granularity. The first level provides global information at multiple resolutions. This is analogous to a human mind processing holistic information for face recognition at varying resolutions. Inner and outer facial information are extracted at the second level. Local facial features play an important role in face recognition by human mind. At the third level features are extracted from the local facial regions.

a. First Level of Granularity

In the first level, face granules are generated by applying the Gaussian and Laplacian operators. The Gaussian operator generates a sequence of low pass filtered images by iteratively convolving each of the constituent images with a 2D Gaussian kernel. The resolution and sample density of the image is reduced between successive iterations and therefore the Gaussian kernel operates on a reduced version of the original image in every iteration. Similarly, the Laplacian operator generates a series of band pass images. Let the granules generated by Gaussian and Laplacian operators be represented by F_{Gri} , where i represents the granule number.

b. Second Level of Granularity

To accommodate the observations of Campbell horizontal and vertical granules are generated by dividing the face image F into different regions

Fig1: Horizontal face granules from the second level of granularity

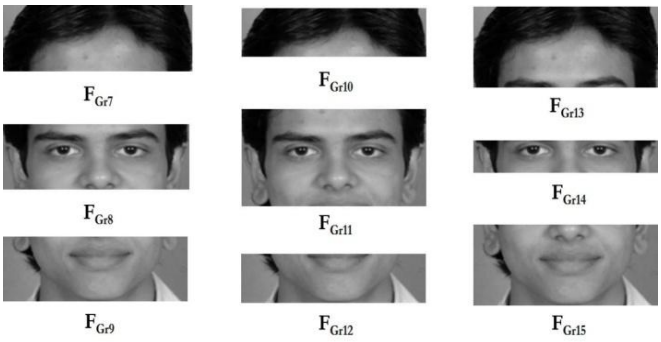


Fig2. Vertical face granules from the second level of granularity

In the above figure $F_{Gr7}$to F_{Gr15} denote the horizontal granules and $F_{Gr16}$to F_{Gr24} denotes the vertical granules.

The second level of granularity provides resilience to variations in inner and outer facial regions. It utilizes the relation between horizontal and vertical granules to address the variations in chin, forehead, ears, and cheeks caused due to plastic surgery procedures.

Feature extraction

In this research, any two (complementing) feature extractors can be used.

They are

a) Extended Uniform Circular Local Binary Patterns (EUCLBP) and

b) Scale Invariant Feature Transform (SIFT) are used.

a. Extended Uniform Circular Local Binary Patterns (EUCLBP):

Extended Uniform Circular Local Binary Pattern (EUCLBP) is a texture based descriptor that encodes exact gray-level differences along with difference of sign between neighboring pixels. For computing EUCLBP descriptor the image is first tessellated into non-overlapping uniform local patches of size 32 × 32. For each local patch, the EUCLBP descriptor is computed based on the 8 neighboring pixels uniformly sampled on a circle (radius=2) centered at the current pixel. The concatenation of descriptors from each local patch constitutes the image signature. Two EUCLBP descriptors are matched using the weighted χ^2 distance.

b. Scale Invariant Feature Transform (SIFT):

SIFT is a scale and rotation invariant descriptor that generates a compact representation of an image based on the magnitude, orientation, and spatial vicinity of image gradients. SIFT, is a sparse descriptor that is computed around the detected interest points. However, SIFT can also be used in a dense manner where the descriptor is computed around pre-defined interest points. In this research, SIFT descriptor is computed in a dense manner over a set of uniformly distributed non-overlapping local regions of size 32×32. SIFT descriptors computed at the sampled regions are then concatenated to form the image signature. Similar to EUCLBP, weighted χ^2 distance is used to compare two SIFT descriptors.

MOEA for Selection of Feature Extractor and Weight Optimization using Genetic algorithm

Feature selection problem embroils around two objectives: 1) select an optimal feature extractor for each granule, and 2) assign proper weight for each face granule. The problem of finding optimal feature extractor and weight for each granule involves searching very large space and finding several suboptimal solutions. Genetic algorithms (GA) are well proven in searching very large spaces to quickly converge to the near optimal solution Therefore; a multi objective genetic algorithm is proposed to incorporate feature selection and weight optimization for each face granule.

Genetic Encoding: A chromosome is a string whose length is equal to the number of face granules i.e. 40 in our case. For simultaneous optimization of two functions, two types of chromosomes are encoded: (i) for selecting feature extractor (referred to as chromosome type1) and (ii) for assigning weights to each face granule (referred to as chromosome type2). Each gene (unit) in chromosome type1 is a binary bit 0 or 1 where 0 represents the SIFT feature extractor and 1 represents the EUCLBP feature extractor. Genes in chromosome type2 have real valued numbers associated with corresponding weights of the 40 face granules. Initial Population: Two generations with 100 chromosomes are populated. One generation has all type1 chromosomes while the other generation has all type2 chromosomes.

1) For selecting feature extractors (type1 chromosome), half of the initial generation (i.e. 50 chromosomes) is et with all the genes (units) as 1, which represents EUCLBP as the feature extractor for

all 40 face granules. The remaining 50 chromosomes in the initial generation have all genes as 0 representing SIFT as the feature extractor for all 40 face granules.

2) For assigning weights to face granules (type2 chromosome), a chromosome with weights proportional to the identification accuracy of individual face granules is used as the seed chromosome. The remaining 99 chromosomes are generated by randomly changing one or more genes in the seed chromosome. Further, the weights are normalized such that the sum of all the weights in a chromosome is 1.

Fitness Function: Both type1 and type2 chromosomes are combined and evaluated simultaneously. Recognition is performed using the feature extractor selected by chromosome type1 and weight encoded by chromosome type2 for each face granule. Identification accuracy, used as the fitness function, is computed on the training set and 10 best performing chromosomes are selected as parents to populate the next generation.

Crossover: A set of uniform crossover operations is performed on parents to populate a new generation of 100 chromosomes. Crossover operation is same for both type1 and type2 chromosomes.

Mutation: After crossover, mutation is performed for type2 chromosomes by changing one or more weights by a factor of its standard deviation in the previous generation. For type1 chromosome, mutation is performed by randomly inverting the genes in the chromosome.

The search process is repeated till convergence and terminated when the identification performance of the chromosomes in new generation do not improve compared to the performance of chromosomes in previous five generations. At this point, the feature extractor and optimal weights for each face granule (i.e. chromosomes giving best recognition accuracy on the training data) are obtained. Genetic optimization also enables discarding redundant and non-discriminating face granules that do not contribute much towards the recognition accuracy (i.e. the weight for that face granule is close to zero). This optimization process leads to both dimensionality reduction and better computational efficiency.

MOEA for Selection of Feature Extractor and Weight Optimization using Firefly Algorithm

In this research, feature selection problem embroils around two objectives: 1) select an optimal feature extractor for each granule, and 2) assign proper weight for each face granule. The problem of finding optimal feature extractor and weight for each granule involves searching very large space and finding several suboptimal solutions.

Firefly Algorithm is well proven in searching very large spaces to quickly converge to the near optimal solution. Therefore, a multi-objective Firefly algorithm is proposed to incorporate feature selection and weight optimization for each face granule.

Multi-objective firefly algorithm (MOFA)

This research learn MOFA algorithm to solve the problem of multi-objective thinking, and its application to the firefly algorithm multi-objective problem solving. Algorithm based on the external feature NP to realize Pareto of multi-objective optimization problem. Below is a basic steps of the algorithm. The population initialization parameters and fireflies position. Initialize external feature NP, and find out all the non dominated Solutions from the initial position of the fireflies filled to the NP.

Combining Face Granules with Multi-objective Evolutionary Learning for Recognition

The granular approach for matching faces altered due to plastic surgery is summarized below.

1) For a given gallery-probe pair, 40 face granules are extracted from each image.

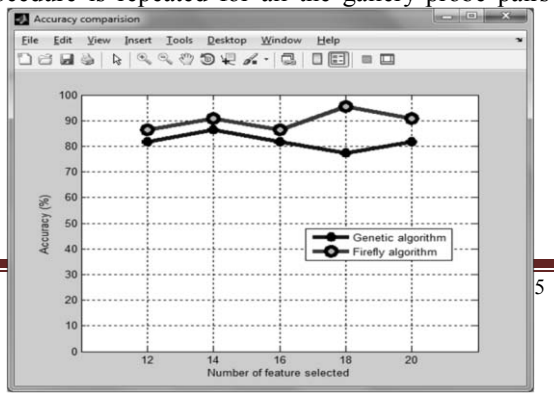
2) EUCLBP or SIFT features are computed for each face granule according to the evolutionary model learned using the training data.

3) The descriptors extracted from gallery and probe images are matched using weighted χ^2 distance measure.

$$\chi^2(a,b) = \sum_{i,j} \omega_j \frac{(a_{i,j} - b_{i,j})^2}{a_{i,j} + b_{i,j}}$$

where a and b are the descriptors computed from face granules pertaining to a gallery-probe pair, i and j correspond to the ith bin of the jth face granule and ω j is the weight of the jth face granule. Here, the weights of each face granule are learnt using the Firefly algorithm.

4) In identification mode (1: N), this procedure is repeated for all the gallery-probe pairs



and top matches are obtained based on the match scores.

3.RESULTS AND DISCUSSION

Plastic Surgery Face Database

One of the major challenge in this research is to prepare a database that contains images of individuals before and after facial plastic surgery. There are several concerns in collecting the database as patients are hesitant in sharing their images. Apart from the issues related to privacy, many who have undergone a disease correcting facial surgery would like to be discrete. To the best of our knowledge, there is no publically available facial plastic surgery database that can be used to evaluate current face recognition algorithms or develop a new algorithm. However, to conduct a scientific experimental study and to analyze the effect of both local and global plastic surgery on face recognition, it is imperative to collect face images before and after plastic surgery.

The plastic surgery face database is a real world database that contains 1800 pre and post surgery images pertaining to 900 subjects. For each individual, there are two frontal face images with proper illumination and neutral expression: the first is taken before surgery and the second is taken after surgery. The database contains 519 image pairs corresponding to local surgeries and 381 cases of global surgery (e.g., skin peeling and face lift). For each individual, there are two frontal face images with proper illumination and neutral expression: the first is taken before surgery and the second is taken after surgery. The database contains 519 image pairs corresponding to local surgeries and 381 cases of global surgery (e.g., skin peeling and face lift). **T**he following graph shows the Accuracy comparison for Genetic algorithm and Firefly algorithm

Comparision graph for GA and Firefly algorithm

Thus the above graph depicts that the proposed Multi objective Firefly algorithm gives out better accuracy than existing Genetic algorithm**.**

4.CONCLUSION AND FUTURE WORK

CONCLUSION

In the present article, Plastic surgery has emerged as a new covariate of face recognition. A novel multi-objective Firefly algorithm called MOFA has been presented. This research presents a multi-objective evolutionary granular algorithm that operates on several granules extracted from a face image. The first level of granularity processes the image with Gaussian and Laplacian operators to assimilate information from multi-resolution image pyramids. The second level of granularity tessellates the image into horizontal and vertical face granules of varying size and information content. The third level of granularity extracts discriminating information from local facial regions. Further, a multi-objective evolutionary Firefly algorithm is proposed for feature selection and weight optimization for each face granule. The evolutionary selection of feature extractor allows switching between two feature extractors (SIFT and EUCLBP) and helps in encoding discriminatory information for each face granule.

FUTURE WORK

Based on the results, we believe that more research is required in order to design an optimal face recognition algorithm that can also account for the challenges due to plastic surgery. It is our assertion that the results of this work would inspire further research in this important area. One possible future research direction would be to use thermal-infrared imagery and compute the thermal differences between pre and post surgery images. However, such an approach first requires creating a large face database that contains pre and post operative thermal infrared images.

REFERENCE

[1] R. Singh, M. Vatsa, H.S. Bhatt, S. Bharadwaj, A. Noore, and S.S. Nooreyezdan, “Plastic surgery: A new dimension to face recognition”, IEEE Transactions on Information Forensics and Security, vol. 5, no. 3, pp. 441–448, 2010.

[2] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, “Robust face recognition after plastic surgery using local region analysis”, in Proceedings of International Conference on Image Analysis and Recognition, 2011, vol. 6754, pp. 191–200.

[3] G. Aggarwal, S. Biswas, P.J. Flynn, and K.W. Bowyer, “A sparse representation approach to face matching across plastic surgery”, in Proceedings of Workshop on the Applications of Computer Vision, 2012, pp. 1–7.

[4] H.S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, “On matching sketches with digital face images”, in Proceedings of International Conference on Biometrics: Theory Applications and Systems, 2010, pp. 1–7.

[5] D.G. Lowe, “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[6] B. Heisele, P. Ho, J. Wu, and T. Poggio, “Face recognition: component based versus global approaches”, Computer Vision and Image Understanding, vol. 91, pp. 6–21, 2003..

[7] B. Gkberk, M. O. Irfanoglu, L. Akarun, and E. Alpaydin, “Learning the best subset of local

features for face recognition", Pattern Recognition, vol. 40, pp. 1520–1532, 2007.

[8] J. Fieldsend and S.Singh, "A Multi-Objective Algorithm based upon Particle Swarm Optimization, an Efficient Data Structure and Turbulence," in Proceedings of UK Workshop on Computational Intelligence (UKCI'02), vol. 2-4, Bermingham, UK, September 2002, pp. 37–44.

[9] C. A. C. Coello, G. T. Pulido, and M. S. Lechuga, "Handling Multiple Objectives With Particle Swarm Optimization," IEEE Transactions on Evolutionary Computation, vol. 8, no. 3, pp. 256–279, June 2004.

[10] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multi-objective Genetic Algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182–197, April 2002.

[11] D. W. Corne, N. R. Jerram, J. D. Knowles, and M. J. Oates, "PESA-II: Region-based Selection in Evolutionary Multiobjective Optimization," in Proceedings of the Genetic and Evolutionary Computing Conference (GECCO-2001). Morgan Kaufmann, 2001, pp. 283–290.

[12] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 10, pp. 1042-1052, 1993.

[13] X. Yi-qiong, L. Bi-cheng and W. Bo, "Face Recognition by Fast Independent Component Analysis and Genetic Algorithm," Proc. of the 4th International Conference on Computer and Information Technology (CIT'04), pp. 194-198, Sept. 2004. .

[14] D.-Q. Dai and H. Yan, "Wavelets and Face Recognition," in Face Recognition, K. Delac and M. Grgic, Eds. I-Tech, Vienna, Austria, 2007, pp.558.

[15] J. Sergent, "Microgenesis of face perception,". In: H.D. Ellis, M.A. Jeeves, F. Newcombe and A. Young, Editors, Aspects of Face Processing, Nijhoff, Dordrecht (1986).

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

DATA STREAM CLASSIFICATION - A SURVEY

Dr. S. Vijayarani[1], Ms. R. Prasannalakshmi[2],

Assistant Professor[1], M.Phil Research Scholar[2],
Department of Computer Science, School of Computer Science and Engineering,
Bharathiar University, Coimbatore, Tamilnadu, India[1,2] .

ABSTRACT

*A*dvanced developments in the hardware and software technologies, computerized data generation and storage have become faster than ever. Data streaming is a challenging task to store, imagine and analyze such rapid large volumes of data. The majority of conventional data mining techniques have to be modified to run in a streaming environment, because of the underlying resource constraint in terms of memory and running time. In addition, the data stream may often show concept drift, because of which adjustment of conventional algorithms becomes extra-challenging. One important conventional data mining problem is that of classification. Classification problem has been extensively studied from a conventional mining perception; it is a great challenging problem in the data stream domain. Data stream classification acting an important role in modern data analysis, where data arrives in a stream and wants to be mined in real time.

Keywords: Data Streams, Classification, Algorithms and Techniques.

I. INTRODUCTION

A data stream is a real-time, continuous, and ordered sequence of items. Some of the examples of data streams are internet traffic streams, stock trading streams, and telephone call streams. Data streams are described as data being unbounded, continuously arriving at a high rate, and being scanned once. With the help of data mining techniques the information and knowledge contained in data streams are extracted. Data stream classification problems have been studied as a major kind of the data analysis tasks in machine learning, statistical inference and data mining. Classification methods characterize the set of supervised learning techniques where a set of dependent variables needs to be predicted based on another set of input attributes. There are two major characteristic approaches under the supervised learning class: classification and regression. Classification is mostly troubled with categorical attributes as dependent variables; still regression is concerned with numerical attributes as its output. The classification process is separated into two phase: model building and model testing. In model building, a learning algorithm runs over a data set to induce a model that could be used to estimate an output. The superiority of this estimation is assessed in the model testing phase.

The typical applications of data stream classification are credit approval, target marketing, medical diagnosis and treatment effectiveness analysis. Classification techniques have involved the attention of researchers due to the importance of their applications. A variety of method such as decision trees, rule based methods, and neural networks are used for the classification problem. Lot of these techniques have been planned to build classification models from static data sets where several passes over the stored data is probable. This is not probable in the case of data streams, in which it is necessary to process the entire data set in one pass. Additionally, the classification problem requirements to be re-designed in the perspective of concept drift which is a unique problem in data streams.

II. DATA STREAM CLASSIFICATION TECHNIQUES

Important data stream classification techniques used in the literature are Ensemble-based Classification, Very Fast Decision Trees (VFDT), On-Demand Classification, On-Line Information Network (OLIN), Lightweight Classification (LW Class), Scalable Classification Algorithm by Learning decision Patterns (SCALLOP) and Adaptive Nearest Neighbor Classification for Data-streams (ANNCAD)[1].

(a) Ensemble-based Classification

This Ensemble based classification approach builds a set of classifiers from the training data. A class label is predicted by using the previously unseen records and by aggregating the predictions made by multiple classifiers. An ensemble classifier builds a

set of 'base classifiers' from the training data [2]. The methods for constructing an ensemble based classification,

- Manipulating training set
- Manipulating input features
- Manipulating class labels
- Manipulating learning algorithms

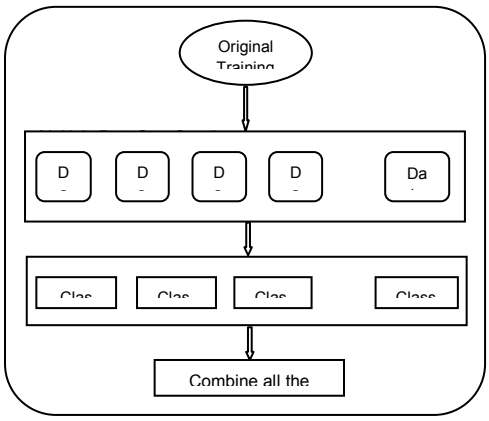


Fig1. Ensemble classification

Manipulating training set: Multiple training sets are created by resembling the data according to some sampling distribution. Sampling distribution determines how likely it is that an example will be selected for training may vary from one trial to another. The classifier is built from each training set using a particular learning algorithm, for example, Bagging & Boosting.

Manipulating input features: Subset of input features chosen to form each training set. Subset can be chosen randomly or based on inputs given by Domain Experts. Good for data that has redundant features. Random Forest is an example which uses DT as its base classifiers.

Manipulating class labels: When number of classes is sufficiently large then the training data is transformed into a binary class problem by randomly partitioning the class labels into 2 disjoint subsets, A0 & A1. Re-labeled examples are used to train a base classifier by repeating the class labeling and model building steps several times, and together of base classifiers is obtained. Example – error correcting output coding.

Manipulating learning algorithm: Learning algorithms can be manipulated in such a way that applying the algorithm several times on the same training data may result in different models. Example – ANN can produce different models by changing network topology or the initial weights of links between neurons**.** Example – ensemble of DTs can be constructed by introducing randomness into the tree growing procedure – instead of choosing the best split attribute at each node, at randomly chooses one of the top k attributes.

First three approaches are generic and these can be applied to any classifier. Fourth approach depends on the type of classifier used. Base classifiers can be generated sequentially or in parallel. Ensemble methods work better with 'unstable classifiers'. Classifiers those are sensitive to minor perturbations in the training set. For Examples, Decision trees, Rule-based, artificial neural networks.

(b) Very Fast Decision Trees (VFDT)

VFDT is very fast, incremental, any time decision tree induction algorithm [4]. The features of VFDT are,

- Use the primary to pick the split at the root
- Sort succeed ones to the leaves
- Pick and choose best attribute
- Continue / Maintain.

Quality of Trees from VFDT

- Make sure most excellent attribute is better than second.
  - That is: $\overline{G}(X_1)-\overline{G}(X_2)>0$
- Using a sample so need bound

$$\overline{G}(X_1)-\overline{G}(X_2)>\varepsilon$$

Complete VFDT System

- Memory management: The memory dominated by sufficient statistics.The deactivates less promising leaves when needed.
- Ties: Wasteful to decide between identical attributes. Check for splits periodically
- Pre-pruning: Only make splits that improve the value of G (.). Early stop on bad attributes. Bootstrap with traditional learner. Rescan dataset when time available. Time changing data streams, Post pruning, Continuous attributes and Batch mode.

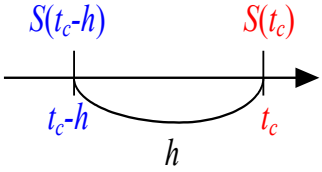
(c) On-Demand Classification

In order to perform effective classification of the stream, it is important to find the correct time-horizon which should be used for classification.

On Demand Classification are constructed by,

- Find the correct time-horizon
- The value of k_{fit}

Find the correct time-horizon



The macro-clusters are created over a user-specified time horizon *h*. Let *S* (*Tc*): the snapshot of micro-clusters at time *Tc* ,*S* (t_c-*h*): the snapshot of micro-clusters at time t_c-*h*. The new sets of micro-clusters N (tc-h) are created by subtracting *S* (t_c-*h*) from *S* (t_c) Subtractive property. Let C_1 and C_2 be two sets of points such that,

$$\Delta\overline{G} < \varepsilon < \tau\, C_1 \supseteq C_2$$

Training Data Stream

A small part of the stream is used for the process of horizon fitting stream segment, K_{fit}: the number of points in the data used and the value small as 1% of the data remaining portion of the training stream is used for the creation and maintenance of the class-specific micro-clusters. ***The value of k $_{fit}$ is***,

- Horizon determined classification accuracy
- Process executed periodically for changes
- k_{fit} should be small enough so that the points in it reflect the immediate locality of t_c
- Q_{fit} :pre-specified number of time units
    - A part of the training stream and the class labels are known a-priori.
- Nearest neighbor procedure (XεQ_{fit})
    - Find the closest micro cluster within N (*Tc*, h) to X. Compare the class label and true label is valuable.

(d) Online Information Network (OLIN)

The system rebuilds the classification model with the most up to date examples. By using the error rate as a guide to concept drift, the frequency of model construction and the window size is adjusted over time. The system applies info-fuzzy techniques for building a tree-like classification model. It uses information theory to calculate the window size. The key idea behind the system is to change the sliding window of the model reconstruction according to the classification error rate. If the model is constant, the window size increases. Thus the occurrence of model building decreases. The info-fuzzy technique for creating a tree-like classification model is referred to as the Info Fuzzy Network (IFN). The tree is unlike than conventional decision trees in that each level of the tree represents only one attribute except the root node layer. The nodes stand for different values of the attribute. The process of constructing this tree has been term as Information Network (IN).

(e) LWClass Algorithm

Lightweight Classification technique is termed as LWClass. LWClass is based on Algorithm Output Granularity (AOG) [5]. The algorithm output granularity introduced the first resource aware data analysis approach that can handle with changeable data rates according to the available memory and the processing speed. The AOG perform the local data analysis on resource controlled devices that generate or receive streams of information. AOG has three stages of mining, adaptation and knowledge integration. LWClass establish with determining the number of instances that could be resident in memory according to the available space. Once a classified data records arrives, the algorithm searches for the nearest instance already stored in the main memory. This is done using a pre-specified distance threshold. This threshold corresponds to the similarity measure tolerable by the algorithm to consider two or more data records as an entry into a matrix. This matrix is a reviewed version of the original data set. If the algorithm finds a nearest neighbor, it checks the class label. If the class label is equal, it increases the weight for this instance by one, or else it decrements the weight by one. If the weight is decrement down to zero, this entry will be out from the memory conserving the limited memory on streaming applications. The algorithm output granularity is controlled by the distance threshold value and is changing overtime to manage with the high speed of the incoming data elements.

The algorithm procedure could be explained as follows:

1. Each record in the data stream contains attribute values for al, a2 . . . an attributes and the class category.
2. According to the data rate and the available memory, the algorithm output granularity is applied as follows: It calculates the distance between the new record and the stored ones. If the distance is less than a threshold, store the average of these two records and increase the weight for this average as an entry by 1. This is in case that both items have the same class category. If they have different class categories, the weight is decreased by 1 and released from memory if the weight reaches zero.

(f) ANNCAD Algorithm

An incremental classification algorithm named as Adaptive Nearest Neighbor Classification for Data-streams (ANNCAD).The algorithm uses Haar Wavelets Transformation for multi-resolution data representation. A grid-based demonstration at each level is used. The process of classification starts according to the majority of nearest neighbors at finer levels. If the finer levels are not capable to differentiate between the classes with a pre-specified threshold, the levels are used in a

hierarchical way. To address the concept drift problem of the evolving data streams, an exponential fade factor is used to decrease the weight of old data in the classification process [6].The drawback of this technique represented the inability of dealing with unexpected concept drifts as the exponential fade factor takes a while to have its effect feel. In fact, the choice of the exponential fade factor is an inherent flexibility which could lead to over-estimation or under-estimation of the rate of concept drift. Both errors would result in a reduction in accuracy.

III. RELATED WORKS

Moharned Medhat Gaber[1] talk about the hardware and software technologies, computerized data generation and storage has become faster than ever. In addition, the data stream may often show concept drift, because of which adjustment of conventional algorithms becomes extra-challenging. One important conventional data mining problem is that of classification. Classification problem has been extensively studied from a conventional mining perception; it is a great challenging problem in data stream domain.

Arkady Zaslavsky and Shonali Krishnaswamy[2] described the classification problems and it is a major category of the data analysis tasks in machine learning, statistical inference and data mining. Classification methods characterized the set of supervised learning techniques which is a set of dependent variables needs to be predicted based on another set of input attributes.

Wang et al. [3] analyzed the generic framework for mining concept drifting data streams. The framework is ensemble-based classification on the observation that many data stream mining algorithms do not address the issue of concept drift in the growing data. The new idea is based on ensemble classification models such as decision trees using Ripper, Bayesian and others to vote for the classification algorithms to increase the accuracy of the predicted output.

Domingos and Hulten [4] have developed a decision tree as Very Fast Decision Trees (VFDT).It is a decision tree knowledge system based on Hoeffding trees. It divides the tree using the current best attribute taking into consideration that the number of examples used assure the Hoeffding bound. Further new technique has the property that its output is nearly identical to that of a conventional learner. VFDT is a complete version of such a method which can address the research issues of data streams.

Gaber et.al [5] studied Lightweight Classification techniques. It is mostly used on Algorithm Output Granularity. The algorithm output granularity (AOG) approach is used for data analysis and handled the unchallengeable data rates according to the available memory. New techniques are established by determining the number of instance variable could be resident into the memory.

Law et.a1 [6] discussed the incremental classification algorithm for Adaptive Nearest Neighbor Classification for Data-streams. It is used for the transformation of multi-resolution data representation. The proposed process starts with challenging to classic and the data record according to the majority nearest neighbors at finer levels. The drawback of this technique represented in inability of dealing with unexpected concept drifts as the exponential fade factors. In future, the choice of the exponential fade factor is an inherent flexibility which could lead to over-estimation or under-estimation of the rate of concept drift. Both errors would result in a reduction in accuracy.

IV. RESEARCH CHALLENGES

This part addresses the main research issues encountered in the context of stream mining.

High Speed natural world of Data Streams: The inherent characteristic of data streams is high speed. The algorithm should be able to get used to the high speed nature of streaming information. The costof building a classification model is higher than the data rate. Still, it is not possible to scan the data more than once. This is referred to as the *one-pass* constraint.

Unbounded Memory Requirements: Classification techniques have need of data to be resident in memory for building the model. The massive amounts of data streams generated rapidly. This challenge has been addressed using load shedding, sampling and aggregation and creating data synopsis. The memory problem is an important motivation behind many of the developed techniques in the area.

Concept Drifting: Concept drifts modify the classifier results over time. This is change in the underlying data patterns. It is also referred to as data stream evolution**.** The capture of such changes would help in updating the classifier model successfully. The use of an out-of-date model could lead to very low classification accuracy.

Challenges in Distributed Applications: A large number of data stream applications run in mobile environments with limited bandwidth such as sensor networks and handheld devices. Thus data structure representation is an important issue. Subsequently, extracting models and patterns locally from data stream generators or receivers, it is important to convey the data mining output to the user. The user could be a mobile user stationary one getting the results from mobile nodes. This is frequently challenging because of the bandwidth limits in transferring data.

Modeling change of mining results over time: Some cases, the user are not involved in mining data stream results, but how these results are changing over a temporal basis. The classification vary could help in understanding the change in data streams over time.

Interactive Mining environment to satisfy user results: Mining data streams is an extremely application oriented field. For the user should be able to change the classification parameters to serve the special needs of the user under the current context. The fast nature of data streams often creates it more difficult to integrate user-interaction.

Hardware and other Technological Issues: The technological problem of mining data streams is an important one. The hardware issues are of special concerns. Small devices generating data streams are not considered for complex computations. Mentally emulators are used for such tasks and it is a real burden for data stream mining applications which run in resource-constrained situations.

V. CONCLUSION

This paper studied the concepts of data streams classification techniques and research challenges. This study gives how the existing approaches Ensemble-based Classification, Very Fast Decision Trees (VFDT), On-Demand Classification, On-Line Information Network (OLIN), Lightweight Classification (LW Class) and Adaptive Nearest Neighbor Classification for Data-streams (ANNCAD) are used to perform the classification in data streams. In future, new classifications techniques are to be proposed to overcome the drawbacks of the existing approaches.

REFERENCES

[1] Aggarwal C., Han J., Wang J., Yu P. S., (2003) A Framework for Clustering Evolving Data Streams, *Proc. 2003 Int. Con$ on Very Large Data Bases(VLDB'03), Berlin, Germany, Sept. 2003.*

[2]Babcock *B.,* Datar M., and Motwani R. (2003) Load Shedding Techniques for Data Stream Systems (short paper) In *Proc. of the 2003 Workshop on Management and Processing of Data Streams (MPDS 2003).*

[3] Burl M., Fowlkes C., Roden J., Stechert A., and Mukhtar S. (1999), Diamond Eye: A distributed architecture for image data mining, in *SPIEDMKD, Orlando.*

[4]Domingos P. and Hulten G. (2000) Mining High-speed Data Streams. In *Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining.*

[5] Dong G., Han J., Lakshmanan L. V. S., Pei J., Wang H. and Yu P. S. (2003) Online mining of changes from data streams: Research problems and preliminary results, In *Proceedings of the 2003 ACM SIGMOD Workshopon Management and Processing of Data Streams.*

[6] Ferrer-Troyano F. J., Aguilar-Ruiz J. S. and Riquelme J. C. (2004) Discovering Decision Rules from Numerical Data Streams, *ACM Symposiumon Applied Computing*.

[7] Gaber, My M., Zaslavsky, A., and Krishnaswamy, S. (2005) Mining Data Streams: A Review. *ACM SIGMOD Record,* Vol. 34, No. 1, June 2005, ISSN: 0163-5808.

[8] Gaber, My M., Krishnaswamy, S., and Zaslavsky, A., (2005). On-board Mining of Data Streams in Sensor Networks, Accepted as a chapter in the forthcoming book *Advanced Methods of Knowledge Discovery from Complex Data,* (Eds.) Sanghamitra Badhyopadhyay, Ujjwal Maulik, Lawrence Holder and Diane Cook, Springer Verlag, to appear.

[9] Hand D. J. (1999) Statistics and Data Mining: Intersecting Disciplines ACM SIGKDD Explorations.

[10] Hulten G., Spencer L., and Domingos P. (2001) Mining Time-Changing Data Streams. ACM SIGKDD Conference.

[11] Jin R. and Aggarwal G. (2003), Efficient Decision Tree Construction on Streaming Data, in Proceedings of ACM SIGKDD Conference.

[12] Last M. (2002) Online Classification of No stationary Data Streams Intelligent Data Analysis.

[13] Law Y., Zaniolo C. (2005) an Adaptive Nearest Neighbor Classification Algorithm for Data Streams, Proceedings of the 9th European Conference on the Principals and Practice of Knowledge Discovery in Databases, Springer Verlag, Porto, Portugal.

[14] Wang H., Fan W., Yu P. and Han J. (2003) Mining Concept-Drifting Data Streams using Ensemble Classifiers, in the *9th ACM International Conference on Knowledge Discovery and Data Mining (SIGkDD), Washington DC, USA.*

[15] Wang K., Zhou S., Fu A., Yu J. (2003) Mining changes of classification by correspondence tracing. *SIAM International Conference on Data Mining.*



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Cyber security and Advanced web Vulnerability Scanning

Mrs. B. Priyadarshini [MCA] [1], Mr.R.Prashanth [2]

[1]Assistant professor [2] BCA & SS Department

[2] [1] Sri Krishna Arts and Science College,Coimbatore.

ABSTRACT

As the popularity of the web increases and web applications become tools of everyday use, the role of web security has been gaining importance as well. The last years have shown a significant increase in the number of web-based attacks. For example, there has been extensive press coverage of re- cent security incidences involving the loss of sensitive credit card information belonging to millions of customers. Many web application security vulnerabilities result from. generic input validation problems. Examples of such vulnersabilities are SQL injection and Cross-Site Scripting (XSS). Although the majority of web vulnerabilities are easy to understand and to avoid, many web developers are, unfor-tunately, not security-aware. As a result, there exist many web sites on the Internet that are vulnerable. This paper demonstrates how easy it is for attackers to automatically discover and exploit application-level vulner- abilities in a large number of web applications. To this end, we developed It, a generic and modular web vulnerability scanner that, similar to a port scanner, automatically analyzes web sites with the aim of finding exploitable SQL injection and XSS vulnerabilities. Using It, we were able to find many potentially vulnerable web sites. To verify the accuracy of It, we picked one hundred interesting web sites from the potential victim list for further analysis and confirmed exploitable flaws in the identified web pages. Among our victims were well-known global companies and a finance ministry. Of course, we notified the administrators of vulnerable sites about potential security problems.

Index terms :

Introduction – Hacking And Hackers - Typical Web Attacks – Sql Injection - Vulnerability Scanner Applications - Coverage Analysis - Applications – Conclusion

1. INTRODUCTION

The web has become an important part of our lives. Every day we interact with a large number of custom-built web applications that have been implemented using a variety of different technologies. The highly heterogeneous nature of the web with its different implementation languages, encoding standards, browsers and scripting environments makes it difficult for web application developers to properly secure their applications and stay up-to-date with emerging threats and newly discovered attacks. A decade ago, applications were often deployed in closed client-server or stand-alone scenarios. At that time, testing and securing an application was an easier task than to-day, where a web application can be accessed by millions of anonymous Internet users. As more and more security critical applications, such as banking systems, governmental transaction interfaces, and e-commerce platforms, are be-coming directly accessible via the web, the role of web application security and defense has been gaining importance. Many web application security vulnerabilities result from generic input validation problems. Examples of such vulnerabilities are SQL injection and Cross-Site Scripting (XSS).Although the majority of web vulnerabilities are easy to understand and to avoid, many web developers are, unfortunately, not security-aware. As a result, there exist a large number of vulnerable applications and web sites on the web. There are two main approaches to testing software applications for the presence of bugs and vulnerabilities:

- In white-box testing, the source code of the application is analyzed in an attempt to track down defective or vulnerable lines of code. This operation is often integrated into the development process by creating add-on tools for common development environments.

- In black-box testing, the source code is not examined directly. Instead, special input test cases are generated and sent to the application. Then, the results returned by the application are analyzed for unexpected behavior that indicate errors or vulnerabilities.

1.1 The need for automated Web Vulnerability Scanner

Manual vulnerability auditing of all your web applications is complex and time-consuming, since it generally involves processing a large volume of data. It also demands a high-level of expertise and the ability to keep track of considerable volumes of code used in a web application. In addition, hackers are constantly finding new ways to exploit your web application, which would mean that you have to constantly monitor the security communities, and find new vulnerabilities in your web application code before hackers discover them.

Automated vulnerability scanning allows you to focus on the already challenging task of building a web application. An automated web application scanner is always on the lookout for new attack paths that hackers can use to access your web application or the data behind it.

Within minutes, an automated web application scanner can scan your web application, identify all the files accessible from the internet and simulate hacker activity in order to identify vulnerable components.

In addition, an automated vulnerability scanner can also be used to assess the code which makes up a web application, allowing it to identify potential vulnerabilities which might not be obvious from the internet, but still exist in the web application, and can thus still be exploited.

2. HACKING AND HACKERS

Hacking is the practice of modifying the features of a system, in order to accomplish a goal outside of the creator's original purpose. The person who is consistently engaging in hacking activities, and has accepted hacking as a lifestyle and philosophy of their choice, is called a hacker. Computer hacking is the most popular form of hacking nowadays, especially in the field of computer security, but hacking exists in many other forms, such as phone hacking, brain hacking, etc. and it's not limited to either of them.Due to the mass attention given to black hat hackers from the media, the whole hacking term is often mistaken for any security related cyber crime. This damages the reputation of all hackers, and is very cruel and unfair to the law abiding ones of them, from whom the term itself originated. In computer networking, *hacking* is any technical effort to manipulate the normal behavior of network connections and connected systems. A *hacker* is any person engaged in hacking. The term "hacking" historically referred to constructive, clever technical work that was not necessarily related to computer systems. Today, however, hacking and hackers are most commonly associated with malicious programming attacks on the Internet and other networks.

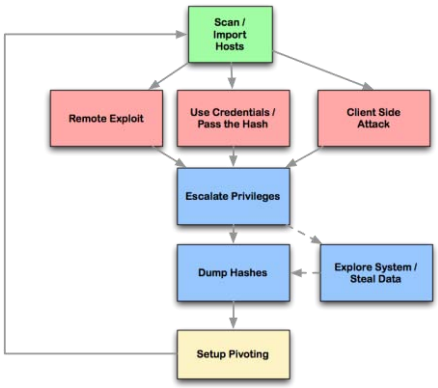


Fig. 1: Hacking – Cyber Security Process

In the computer security context, a hacker is someone who seeks and exploits weaknesses in a computer system or computer network. Hackers may be motivated by a multitude of reasons, such as profit, protest, or challenge. They are subject to the long standing hacker definition controversy about the true meaning of the term hacker. In this controversy, the term hacker is reclaimed by computer who argue that someone breaking into computers is better called a cracker, not making a difference between computer criminals (black hats) and computer security experts (white hats).Some white hat hackers claim that they also deserve the title hacker, and that only black hats should be called crackers.

3. TYPICAL WEB ATTACKS

3.1 SQL INJECTION

SQL Injection is one of the many web attack mechanisms used by hackers to steal data from organizations. It is perhaps one of the most common application layer attack techniques used today. It is the type of attack that takes advantage of improper coding of your web applications that allows hacker to inject SQL commands into say a login form to allow them to gain access to the data held within your database.

In essence, SQL Injection arises because the fields available for user input allow SQL statements to pass through and query the database directly.

SQL injection attacks are based on injecting strings into database queries that alter their intended use. This can occur if a web application does not properly filter (sanitize) user input. There are many

varieties of SQL. Most dialects are loosely based on the most recent ANSI standard SQL-92 [17]. The typical unit of execution in the SQL language is the query, a collection of statements that are aimed at retrieving data from or manipulating records in the database. A query typically results in a single result set that contains the query results. Apart from data retrieval and updates, SQL statements can also modify the structure of databases using Data Definition Language statements ("DDL") A web application is vulnerable to an SQL injection attack if an attacker is able to insert SQL statements into an existing SQL query of the application. This is usually achieved by injecting malicious input into user fields that are used to compose the query. For example, consider a web application that uses a query such as the one shown in Listing 1 for authenticating its users.

This form of SQL injection occurs when user input is not filtered for escape characters and is then passed into a SQL statement. This results in the potential manipulation of the statements performed on the database by the end-user of the application.

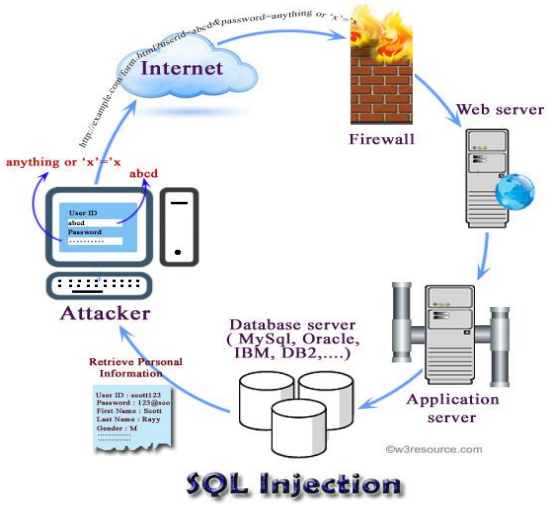


Figure 2: SQL injection process

The following line of code illustrates this vulnerability:

statement = "SELECT * FROM users WHERE name ='" + userName + "';"

This SQL code is designed to pull up the records of the specified username from its table of users. However, if the "userName" variable is crafted in a specific way by a malicious user, the SQL statement may do more than the code author intended. For example, setting the "userName" variable as:

' or '1'='1

or using comments to even block the rest of the query (there are three types of SQL comments):[11]

' or '1'='1' -- '

' or '1'='1' ({ '

' or '1'='1' /* '

renders one of the following SQL statements by the parent language:

SELECT * FROM users WHERE name = '' OR '1'='1';

SELECT * FROM users WHERE name = '' OR '1'='1' -- ';

If this code were to be used in an authentication procedure then this example could be used to force the selection of a valid username because the evaluation of '1'='1' is always true.

The following value of "userName" in the statement below would cause the deletion of the "users" table as well as the selection of all data from the "userinfo" table (in essence revealing the information of every user), using an API that allows multiple statements:

a';DROP TABLE users; SELECT * FROM userinfo WHERE 't' = 't

This input renders the final SQL statement as follows and specified:

SELECT * FROM users WHERE name = 'a';DROP TABLE users; SELECT * FROM userinfo WHERE 't' = 't';

While most SQL server implementations allow multiple statements to be executed with one call in this way, some SQL APIs such as PHP's mysql_query(); function do not allow this for security reasons. This prevents attackers from injecting entirely separate queries, but doesn't stop them from modifying queries.

4. VULNERABILITY SCANNER APPLICATIONS

In the present study we have tested four commercial web vulnerability scanners widely used, including two different versions of a specific brand. The following paragraphs briefly introduce these scanners.HP WebInspect "*performs web application security testing and assessment for today's complex web applications, built on*

emerging Web 2.0 technologies. HP Web Inspect delivers fast scanning capabilities, broad security assessment coverage and accurate web application security scanning results" [10]. This tool includes pioneering assessment technology, including simultaneous crawl and audit (SCA) and concurrent application scanning. It is a broad application that can be applied for penetration testing in web-based applications. IBM Rational AppScan "*is a leading suite of automated Web application security and compliance assessment tools that scan for common application vulnerabilities*. This tool is suitable for users ranging from non-security experts to advanced users that can develop extensions for customized scanning environments. IBM AppScan can be used for security testing in web applications, including web services. Acunetix Web Vulnerability Scanner "*is an automated web application security testing tool that audits your web applications by checking for exploitable hacking vulnerabilities*" .Besides web services, Web Vulnerability Scanner can be applied for security testing in web applications in general.

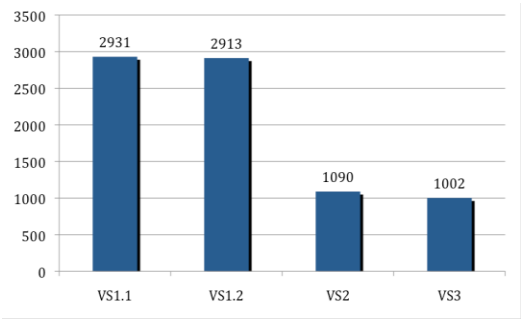


Figure 3. Application errors detected.

4.1 Coverage analysis

A key aspect is to understand the coverage of the vulnerabilities detected. Coverage compares the number of vulnerabilities detected against the total number of vulnerabilities. Obviously, in our case it is impossible to know how many vulnerabilities were not disclosed by any of the scanners (we do not have access to the source code). Thus, it is not possible to calculate the coverage. However, it is still possible to make a relative comparison based on the data available. In practice, we know the total number of vulnerabilities detected (which correspond to the union of the vulnerabilities detected by the four scanners after removing the false positives) and the number of vulnerabilities detected by each individual scanner. Based on this information it is possible to get an optimistic coverage indicator for each scanner (i.e., the real coverage will be lower than the value presented). Obviously, this is relevant only for SQL Injection vulnerabilities as it is the only type that is detected by all the scanners. Table 3 presents the coverage results. As shown, 149 different SQL Injection vulnerabilities were detected (as before, we decided to include the doubtful situations as existing vulnerabilities). Each scanner detected a subgroup of these vulnerabilities, resulting in partial detection coverage. VS1.1 and VS1.2 present quite good results. On the other hand, the coverage of VS2 and VS3 is very low.

Scanner	# SQL Injection Vuln.	Coverage %
VS1.1	130	87.2%
VS1.2	142	95.3%
VS2	25	16.8%
VS3	26	17.4%
Total	**149**	**100.0%**

Table : 1 Coverage of SQL injection

The results presented before allowed us to observe some interesting aspects. The first observation is that different scanners detected different types of faults. SQL Injection was the only type that was detected by all scanners. The two scanners of the same brand (VS1.1 and VS1.2) were the only ones that detected XPath and code execution vulnerabilities. Only one scanner (VS3) detected vulnerabilities related to buffer overflow, username and password disclosure, and server path disclosure. VS2 only detected SQL Injection vulnerabilities.SQL Injection vulnerabilities are the dominant type in the web services tested (see Figure 4). However, different scanners detected different vulnerabilities of this type. In fact, VS1.1 and VS1.2 detected a huge number of vulnerabilities (215 and 225 respectively) while VS2 and VS3 detected a very low number (25 and 35 respectively). A key observation was a very large number of false
positives. In fact, for three of the scanners the percentage of false positives was more than 25%. VS2 presented zero false positives, but 8 out of the 25 SQL Injection vulnerabilities detected by this scanner remained as doubtful (i.e., could not be manually confirmed as real vulnerabilities nor as false positives).
A very low coverage, lower than 18%, was observed for two of the scanners (VS2 and VS3), while the other two scanners (VS1.1 and VS1.2) present a coverage superior to 87%. Note that this value represents an optimistic coverage, as the real coverage of the tested scanners (at least for the 300 web services used in the experiments) is definitely lower than the value observed.

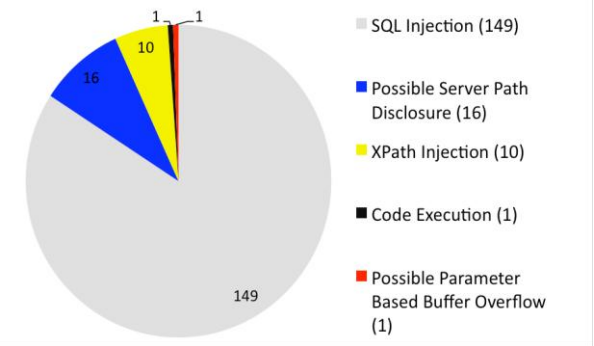


Figure 4: Vulnerabilities distributed per type.

5. CONCLUSION:

Many web application security vulnerabilities result from generic input validation problems. Examples of such vulnerabilities are SQL Injection and Cross-Site Scripting (XSS). Although the majority of web vulnerabilities are easy to understand and avoid, many web developers are, unfortunately, not security-aware and there is general consensus that there exist a large number of vulnerable applications and web sites on the web. The main contribution of this paper is to show how easy it is for attackers to automatically discover and exploit application-
level vulnerabilities in a large number of web applications. To this end, we presented It, a generic and modular web vulnerability scanner that analyzes web sites for exploitable SQL and XSS vulnerabilities. We used It to identify a large number of potentially vulnerable web sites. Moreover, we selected one hundred of these web sites for further analysis and manually confirmed exploitable flaws in the identified web pages. Among our victims were well-known global companies, computer security organizations, and governmental and educational institutions. We believe that it is only a matter of time before attackers start using automated vulnerability scanning tools such as It to discover web vulnerabilities that they can exploit. Such vulnerabilities, for example, could be used to launch phishing attacks that are difficult to identify even by technically more sophisticated users. With this paper, we hope to raise awareness and provide a tool available to web site administrators and web developers to proactively audit the security of their applications. Thus, the advanced vulnerability web scanner is one of the best tool thing to detect the vulnerabilities online.

REFERENCES:

[1] CERT. Advisory CA-2000-02: malicious HTML tags embedded in client web requests. http://www.cert.org/advisories/CA-2000-02.html ,2000.

[2] W3C World Wide Web Consortium. HTTP - Hypertext Transfer Protocol. http://www.w3.org/Protocols/ , 2000.

[3] Microsoft Corporation. Architecture and Design Review for Security. http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnnets%ec/html/THCMCh05.asp, 2005.

[4] Microsoft Corporation. ISAPI Server Extensions and Filters. http://msdn.microsoft.com/library/default.asp?url=/library/en-us/vccore%98/HTML/_core_isapi_server_extensions_and_filters.asp, 2005.

[5] Microsoft Corporation. Microsoft .NET Framework Development Center. http://msdn.microsoft.com/netframework/, 2005.

[6] Microsoft Corporation. System.Reflection Namespace. http://msdn.microsoft.com/library/default.asp?url=/library/en-us/cpref/%html/frlrfsystemreflection.asp, 2005.

[7] David Cruwys. C Sharp/VB - Automated WebSpider / WebRobot. http://www.codeproject.com/csharp/DavWebSpider.asp, March 2004.

[8] Stuttard, D., Pinto, M., "The Web Application Hacker's Handbook: Discovering and Exploiting Security Flaws", Wiley, ISBN-10: 0470170778, October 2007

[9] Ounce, http://www.ouncelabs.com/

[10] Pixy, http://pixybox.seclab.tuwien.ac.at/pixy/

[11] Acunetix Web Vulnerability Scanner, 2008, http://www.acunetix.com/vulnerability-scanner/

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Manhattan Grid & Random Way Mobility based on Performance Analysis of MANET Routing Protocols

[1]G.Kalpana, [2]A.Preethi

[1]Assistant Professor, [2]Research Scholar,
Department of Computer Science , Sri Ramakrishna College of Arts and Science for women
Coimbatore-44

ABSTRACT

Mobile ad hoc networks (MANETs) are a set of mobile nodes which are self-configuring and connected by wireless links routinely used to evaluate and manage trust relationships between network nodes and enhance network security. Due to the lack of a central power, securitizing the routing process becomes a difficult task however trust management frameworks themselves are vulnerable to attacks results in performance degradation. In wireless network technology, simulative analysis is a significant method to understand the performance of routing protocol. The most widely used ad-hoc routing protocols are Ad-hoc On-Demand Distance Vector Routing (AODV), Destination Sequenced Distance Vector (DSDV) and Dynamic Source Routing (DSR). In this two parameters were simulated using Manhattan Grid and Random Way Mobility Model.

Index Terms - MANET, AODV, DSDV, DSR, Manhattan Grid Mobility Model, Random Way Mobility Model.

I. INTRODUCTION

Ad hoc networks are multi-hop wireless networks dynamically constructed by mobile nodes without the aid of any established infrastructure. In ad hoc networks, nodes have to cooperate with each other to support the network functions due to the lack of intermediate servers. A mobile ad hoc network (MANET) is a self-configuring infrastructure less network of mobile devices connected by wireless. *Ad hoc* is Latin and means "for this purpose". Each device in a MANET is free to move independently in any direction, and will therefore change its links to other devices frequently. Each must forward traffic unrelated to its own use, and therefore be a router. The primary challenge in building a MANET is equipping each device to continuously maintain the information required to properly route traffic. Such networks may operate by themselves or may be connected to the larger Internet. Each node must forward traffic unrelated to its own use, and therefore be a router. The primary challenge in building a MANET is equipping each device to continuously maintain the information required to properly route traffic. Therefore, nodes are required to relay packets on behalf of other nodes in order to deliver data across the network.

The characteristics of MANET like dynamic topology, lack of fixed infrastructure, vulnerability of nodes and communication channel, lack of traffic concentration points, limited power, computational capacity, memory, and bandwidth that make the task of achieving a secure and reliable communication that more difficult to use for applications like on-line banking, business sensitive applications, and transfers of military information.

The main aim of this paper is to use the mobility models and to analyze the performance of different routing protocols under mobility. The paper is organized as follows: section 2 discusses the major mobile ad-hoc routing protocols used in this evaluation study, section 3 presents the Random way and Manhattan Grid Mobility Model, section 4 discusses the simulation results, section 5 presents the conclusion.

II.ROUTING PROTOCOLS

Routing protocols in mobile networks are subdivided into two basic classes.

- Proactive routing protocols
- Reactive routing protocols

The proactive routing protocols are also called table-driven protocols it maintains the routing information of the all participant nodes and update their routing information frequently irrespective of the routing request. The Reactive routing protocols create and maintain routes only if these are needed, on demand. It uses connection establishment process for communication. These protocols usually use distance-vector routing algorithms that keep only information about next hops to

adjacent neighbours and costs for paths to all known destinations. Three routing protocols are considered in this paper namely, AODV, DSR and DSDV.

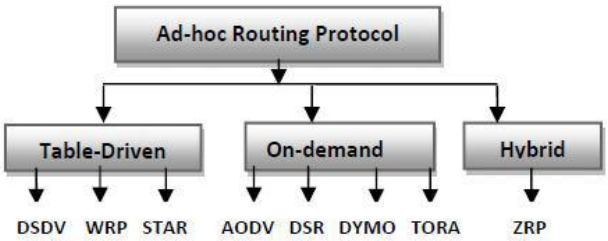


Figure 1.Ad-hoc routing protocols

III. OVERVIEW OF DIFFERENT ROUTING PROTOCOLS

A. AD-HOC ON-DEMAND DISTANCE VECTOR ROUTING (AODV)

The Ad Hoc on-Demand Distance Vector Routing (AODV) [8] is a routing protocol for mobile ad hoc networks (MANETs) and other wireless ad-hoc networks provides on-demand route discovery. It is a reactive routing protocol, meaning that it establishes a route to a destination only on demand. Whenever the nodes need to send data to the destination, if the source node doesn't have routing information in its table, route discovery process begins to find the routes from source to destination. AODV maintains a routing table. Each node has a sequence number. When a node wants to initiate route discovery process, it includes its sequence number and the most fresh sequence number it has for destination. The intermediate node that receive the RREQ packet, replay to the RREP packet only when the sequence number of its path is larger than or identical to the sequence number comprised in the RREQ packet. A reverse path from the intermediate node to the source forms with storing the node's address from which initial copy of RREQ. Thus, at the end of this request-response cycle a bidirectional route is established between the requesting node and the destination. When a node loses connectivity to its next hop, the node invalidates its route by sending an RERR to all nodes that potentially received its RREP.

B. DESTINATION-SEQUENCED DISTANCE VECTOR-DSDV

DSDV belongs to the class of proactive routing protocols. DSDV has the feature of the distance-vector protocol in that each node holds a routing table including the next-hop information for each possible destination. Each entry has a sequence number. If a new entry is obtained, the protocol prefers to select the entry having the largest sequence number. If their sequence number is the same, the protocol selects the metric with the lowest value. In this protocol, packets are routed between nodes of an ad hoc network using routing tables stored at each node. Each routing table, at each node, contains a list of the addresses of every other node in the network. Along with each node's address, the table contains the address of the next hop for a packet to take in order to reach the node. This protocol motivated for the use of data exchange along changing and arbitrary paths of interconnection which may not be close to any base station.

C. DYNAMIC SOURCE ROUTING (DSR)

Dynamic Source Routing (DSR), belongs to the class of reactive protocols and allows to dynamically discover a route across multiple network hops to any destination. DSR uses no periodic routing of messages thereby reducing network bandwidth overhead, conserving battery power and avoiding large routing updates throughout the ad-hoc network. There are two main components, called Route Discovery and Route Maintenance. Route Discovery is the mechanism by which a node wishing to send a packet to a destination obtains a path to the destination. Route Maintenance is the mechanism by which a node detects a break in its source route and obtains a correct route. The sender knows the complete hop by hop route to the destination.

IV. SIMULATION ENVIRONMENT

A. SIMULATION MODEL

In this paper network simulations have been carried out using Network Simulator version 2 (NS-2.34) of AODV, DSR and DSDV and its associated tools for analysis of results.

B. SIMULATION PARAMETERS

Table 1.1: Simulation Parameters

Characteristics	**Manhattan Grid mobility**	**Random Way mobility**
Simulator	NS-2(version 2.34)	OPNET 14.5
Channel type	Wireless Channel	Wireless Channel
Protocols	AODV,DSR and DSDV	AODV,DSR and DSDV
Simulation Duration	100s	100s
Number of Nodes	10-50	10-50
Movement model	Manhattan Grid Model	Random Way Model
MAC Layer Protocol	802.11	802.11

C. *PEFORMANCE PARAMETERS*

This paper analysed the following important performance parameters for comparing the AODV, DSR and DSDV routing protocols.

1. Packet Delivery Fraction: It is the ratio of all received data packets successfully at destinations and all data packets sent by sources.
2. Average end-to-end Delay: It represents the delay encountered between the sending and receiving of the packets. Delay indicates how long it the packet takes to travel from source to destination. This caused by buffering during route discovery, retransmission delay and transfer times of data packet.

D. *SIMULATION RESULTS*

In this paper, an attempt was made to compare all the three protocols under the random way mobility and manhattan grid scenario.

PACKET DELIVERY FRACTION

Packet delivery fraction in manhattan simulator and random way mobility, 10 nodes were randomly placed. While comparing AODV, DSR and DSDV, the performance of AODV is better.

AVERAGE END TO END DELAY

Average end-to-end delay in manhattan simulator and random way mobility, 10 nodes were randomly placed. While comparing AODV, DSR and DSDV, the performance of AODV and DSR is better.

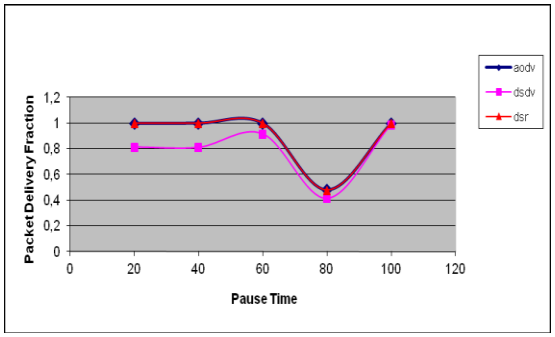


Fig :1 Packet Delivery Fraction in Manhattan

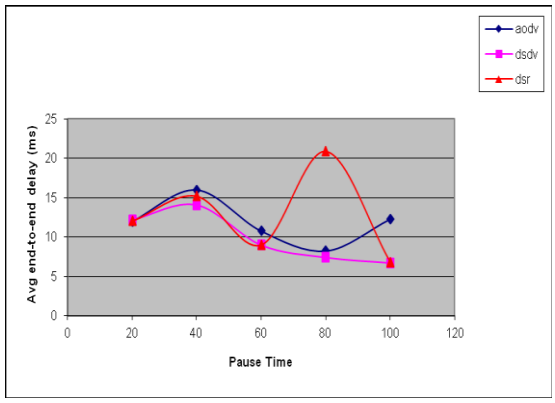


Fig: 2 Average End-End Delay in Manhattan

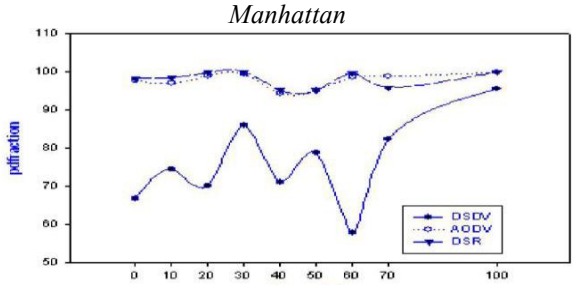


Fig.3 Pause Time vs pdfraction (Fixes 10 nodes)

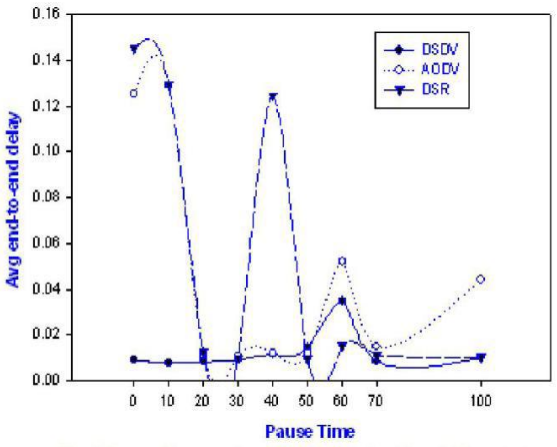


Fig.4 Pause Time vs Avg End-End delay(Fixed 10 nodes)

V. CONCLUSION

In this paper, only two performance measures i.e. Packet Delivery Fraction and Average end-to-end delay were evaluated under the Random Way Mobility and Manhattan Grid Mobility Model. These performance were analyzed using three routing protocols i.e. AODV, DSR and DSDV. In Random Way Model, AODV performs better than DSR and DSDV. Hence, AODV protocol can be chosen in this type of mobility model. In Manhattan Grid Mobility Model, AODV and DSR experience the highest Packet Delivery Fraction and end-

to-end Delay. And also AODV has a slightly higher average end-to-end Delay than DSR.

VI. FUTURE SCOPE

Hybrid routing protocol can be simulated and tested under manhattan and random way mobility model and the result can be determined.

REFERENCES

[1] M. K. Jeya Kumar, R.S. Rajesh, "A Survey of MANET Routing Protocols in Mobility Models", Int. Journal of Soft Computing 4(3), pp. 136-141, 2009.

[2] D.B. Johnson and D.A. Maltz, "Dynamic source routing in ad hoc wireless networks", Mobile Computing (Kluwer Academic, 1996) chapter 5, pp. 153–181.

[3] M.S. Murty and M.V. Das, "Performance evalution of MANET routing protocols using reference point group mobility and random waypoint models", International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC), Vol.2, No.1, March 2011.

[4] M.Sreerama Murty and M.Venkat Das, "Performance Evalution of MANET Routing Protocols using Reference Point Group Mobility and Random Way Point Models", International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.2, No.1, March 2011.

[5] N.Aschenbruck, E.Gerhands-Padilla, P.Martini, "A Surveyon mobility models for Performance analysis in TacticalMobile networks," Journal of Telecommunication andInformation Technology, Vol.2 pp.54-61,2008.

[6] Bai, Fan,Helmy, Ahmed (2006)"A Survey of Mobility Models in Wireless Adhoc Networks." (Chapter 1 in Wireless Ad-Hoc Networks. Kluwer Academic. 2006).

[7] Yun Ge, Guojun Wang and Jie Wu, "Node-Disjoint Multipath Routing with Group Mobility in MANETs", Networking, Architecture and Storage (NAS), 2010 IEEE Fifth International Conference on, pp. 18-189, 2010.

[8] G.S. Malkin and M.E. Steenstrup, Distancevector routing, in: Routing in Communications Networks, ed. M.E. Steenstrup (Prentice Hall,1995) chapter 3, pp. 83–98.

[9] D.B. Johnson and D.A. Maltz, Dynamic source routing in ad hoc wireless etworks, in: Mobile Computing (Kluwer Academic, 1996) chapter 5, pp. 153–181.

[10] J. Broch, D.B. Johnson and D.A. Maltz, The dynamic source routing protocol for mobile ad hoc networks, IETF MANET Working Group, Internet-Draft (October 1999).

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Mobile Ad Hoc Networks Applications and Techniques

[1]Mrs.S.Rajanandini, [2]Ms.N.Vikasini,

[1]Assistant professor, BCA & S.S Department,

[2] II M.Sc SS, BCA & S.S Department,

Sri Krishna Arts and Science College,Coimbatore,

ABSTRACT

In the past few years, we have seen a rapid expansion in the field of mobile computing due to the proliferation of inexpensive, widely available wireless devices. However, current devices, applications and protocols are solely focused on cellular or wireless local area networks (WLANs), not taking into account the great potential offered by mobile ad hoc networking. A mobile ad hoc network is an autonomous collection of mobile devices (laptops, smart phones, sensors, etc.) that communicate with each other over wireless links and cooperate in a distributed manner in order to provide the necessary network functionality in the absence of a fixed infrastructure. This type of network, operating as a stand alone network or with one or multiple points of attachment to cellular networks or the Internet, paves the way for numerous new and exciting applications. Application scenarios include, but are not limited to: emergency and rescue operations, conference or campus settings, car networks, personal networking, etc.This paper provides insight into the potential applications of ad hoc networks and discusses the technological challenges that protocol designers and network developers are faced with. These challenges include routing, service and resource discovery,Internet connectivity, billing and security.

1. INTRODUCTION

A mobile ad hoc network (MANET) is a self-configuring infrastructureless network of mobile devices connected by wireless. *Ad hoc* is Latin and means "for this purpose".[1]

Each device in a MANET is free to move independently in any direction, and will therefore change its links to other devices frequently. Each must forward traffic unrelated to its own use, and therefore be a router. The primary challenge in building a MANET is equipping each device to continuously maintain the information required to properly route traffic. Such networks may operate by themselves or may be connected to the larger Internet.MANETs are a kind of Wireless ad hoc network that usually has a routable networking environment on top of a Link Layer ad hoc network.

The growth of laptops and 802.11/Wi-Fi wireless networking have made MANETs a popular research topic since the mid-1990s. Many academic papers evaluate protocols and their abilities, assuming varying degrees of mobility within a bounded space, usually with all nodes within a few hops of each other. Different protocols are then evaluated based on measures such as the packet drop rate, the overhead introduced by the routing protocol, end-to-end packet delays, network throughput etc.

2. HOW AD HOC NETWORKS WORK

Ad Hoc Networks are useful in areas that have no fixed infrastructure and hence need alternative ways to deliver services. Ad Hoc Networks work by having mobile devices connect to each other in the transmission range through automatic configuration, i.e., setting up an ad hoc network that is very flexible. In other words there is no intervention of any controller that goes ahead and gathers data from all nodes and organizes it. All data gathering and cross-node data transfer is taken care of by the nodes themselves.

Ad Hoc Networks are a major goal towards the evolution of 4G (Fourth generation) devices. In the nodes of the Ad Hoc Networks, computing power

and network connectivity are embedded in virtually every device to bring computation to users, no matter

Cellular network	Ad Hoc Network
Fixed, pre-located cell sites and base stations.	No fixed base stations, very rapid deployment.
Static backbone network topology.	Highly dynamic network topologies with multi-hop.
Relatively benign environment and stable connectivity.	Hostile environment (losses, noise) and sporadic connectivity.
Detailed planning before base stations can be installed.	Ad hoc network automatically forms and adapts to changes.

where they are, or under what circumstances they work. These devices personalize themselves to find the information or software they need. The strife is to make use of all technologies available without making any major change to the user's behavior. There is also work going on to make the seamless integration of various networks possible, i.e., integration of LAN, WAN, PAN and Ad Hoc Networks. But there is still a lot of work to be done to make this completely possible.

Node mobility in an ad hoc network causes frequent changes of the network topology. Figure 1 shows such an example: initially, nodes A and D have a direct link between them. When D moves out of A's radio range, the link is broken. However, the network

is still connected, because A can reach D through C, E, and F.

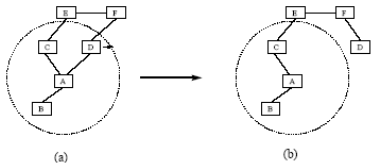


Figure 1: Topology change in ad hoc networks: nodes A, B, C, D, E, and F constitute an ad hoc network. The circle represents the radio range of node A. The network initially has the topology in (a). When node D moves out of the radio range of A, the network topology changes to the one in (b).

Spread Spectrum Techniques are used in the implementation of Ad Hoc Networks because spread spectrum helps to reduce interference from other sources. Also it helps in bandwidth reuse. The boundaries of Ad Hoc Networks are not absolute and hence it is possible that when certain nodes stray into the area of influence of certain transmitters these may get affected by their signals. The use of Spread Spectrum (SS) makes sure that this does not happen as the spreading code and the de-spreading code should ideally be the same. This same technique provides the method for frequency reuse.

***Differences between Cellular Network and Ad Hoc Network*:**

3. BASIC OPERATION

3.1. Overview

To send a packet to another host, the sender constructs a *source route* in the packet's header, giving the
address of each host in the network through which the packet should be forwarded in order to reach the destination host. The sender then transmits the packet over its wireless network interface to the first hop identified in the source route. When a host receives a packet, if this host is not the final destination of
the packet, it simply transmits the packet to the next hop identified in the source route in the packet's header. Once the packet reaches its final destination, the packet is delivered to the network layer software on that host.Each mobile host participating in the ad hoc network maintains a *route cache* in which it caches source routes that it has learned. When one host sends a packet to another host, the sender first checks its routecache for a source route to the destination. If a route is found, the sender uses this route to transmit thepacket. If no route is found, the sender may attempt to discover one using the *route discovery* protocol.While waiting for the route discovery to complete, the host may continue normal processing and may sendand receive packets with other hosts. The host may buffer the original packet in order to transmit it once theroute is learned from route discovery, or it may discard the packet, relying on higher-layer protocol softwareto retransmit the packet if needed. Each entry in the route cache has associated with it an expiration period,after which the entry is deleted from the cache.While a host is using any source route, it monitors the continued correct operation of that route. Forexample, if the sender, the destination, or any of the other hosts named as hops along a route move out ofwireless transmission range of the next or previous hop along the route, the route can no longer be used toreach the destination. A route will also no longer work if any of the hosts along the route should fail or bepowered off. This monitoring of the correct operation of a route in use we call *route maintenance*. Whenroute maintenance detects a

problem with a route in use, route discovery may be used again to discover anew, correct route to the destination.This section describes the basic operation of route discovery and route maintenance.

3.2. Route Discovery

Route discovery allows any host in the ad hoc network to dynamically discover a route to any other hostin the ad hoc network, whether directly reachable within wireless transmission range or reachable throughone or more intermediate network hops through other hosts. A host initiating a route discovery broadcasts4a *route request* packet which may be received by those hosts within wireless transmission range of it. Theroute request packet identifies the host, referred to as the *target* of the route discovery, for which the routeis requested. If the route discovery is successful the initiating host receives a *route reply* packet listing asequence of network hops through which it may reach the target.
In addition to the address of the original initiator of the request and the target of the request, each route request packet contains a *route record*, in which is accumulated a record of the sequence of hops taken by theroute request packet as it is propagated through the ad hoc network during this route discovery. Each routerequest packet also contains a unique *request id*, set by the initiator from a locally-maintained sequencenumber. In order to detect duplicate route requests received, each host in the ad hoc network maintains alist of the initiator address, request id_ pairs that it has recently received on any route request.When any host receives a route request packet, it processes the request according to the following steps:1. If the pair initiator address, request id_for this route request is found in this host's list of recentlyseen requests, then discard the route request packet and do not process it further.2. Otherwise, if this host's address is already listed in the route record in the request, then discard theroute request packet and do not process it further.3. Otherwise, if the target of the request matches this host's own address, then the route record in thepacket contains the route by which the request reached this host from the initiator of the route request.Return a copy of this route in a *route reply* packet to the initiator.4. Otherwise, append this host's own address to the route record in the route request packet, and re-broadcast the request.The route request thus propagates through the ad hoc network until it reaches the target host, which thenreplies to the initiator. The original route request packet is received only by those hosts within wirelesstransmission range of the initiating host, and each of these hosts propagates the request if it is not thearget and if the request does not appear to this host to be redundant. Discarding the request because the host's address is already listed in the route record guarantees that no single copy of the request can propagate around a loop. Also discarding the request when the host has recently seen one with the same initiator address, request id removes later copies of the request that arrive at this host by a different route. In order to return the *route reply* packet to the initiator of the route discovery, the target host must have aroute to the initiator. If the target has an entry for this destination in its route cache, then it may send the routereply packet using this route in the sameway as is used in sending any other packet (Section 3.1). Otherwise,the target may reverse the route in the route record from the route request packet, and use this route to sendthe route reply packet. This, however, requires the wireless network communication between each of thesepairs of hosts to work equally well in both directions, which may not be true in some environments or withsome MAC-level protocols. An alternative approach, and the one we have currently adopted, is for thishost to piggyback the route reply packet on a route request targeted at the initiator of the route discovery towhich it is replying.

3.3. Route Maintenance

Conventional routing protocols integrate route discovery with route maintenance by continuously sendingperiodic routing updates. If the status of a link or router changes, the periodic updates will eventually reflect the changes to all other routers, presumably resulting in the computation of new routes. However, using route discovery, there are no periodic messages of any kind from any of the mobile hosts. Instead, while a route is in use, the route maintenance procedure monitors the operation of the route and informs the senderof any routing errors.
Since wireless networks are inherently less reliable than wired networks [1], many wireless networks utilize a hop-by-hop acknowledgement at the data link level in order to provide early detection and retransmissionof lost or corrupted packets. In these networks, route maintenance can be easily provided, since ateach hop, the host transmitting the packet for that hop can determine if that hop of the route is still working.If the data link level reports a transmission problem for which it cannot recover (for example, because themaximum number of retransmissions it is willing to attempt has been exceeded), this host sends a *routeerror* packet to the original sender of the packet encountering the error. The route error packet contains theaddresses of the hosts at both ends of the hop in error: the host that detected the error and the

host to which itwas attempting to transmit the packet on this hop. When a route error packet is received, the hop in error isremoved from this host's route cache, and all routes which contain this hop must be truncated at that point.If the wireless network does not support such lower-level acknowledgements, an equivalent acknowledgementsignal may be available in many environments. After sending a packet to the next hop mobilehost, the sender may be able to hear that host transmitting the packet again, on its way further along thepath, if it can operate its wireless network interface in promiscuous mode. For example, in Figure 1, host*A* may be able to hear *B*'s transmission of the packet on to *C*. This type of acknowledgement is knownas a *passive acknowledgement* [11]. In addition, existing transport or application level replies or acknowledgementsfrom the original destination could also be used as an acknowledgement that the route (or thathop of the route) is still working. As a last resort, a bit in the packet header could be included to allow a host transmitting a packet to request an explicit acknowledgement from the next-hop receiver. If no otheracknowledgement signal has been received in some time from the next hop on some route, the host coulduse this bit to inexpensively probe the status of this hop on the route.As with the return of a route reply packet, a host must have a route to the sender of the original packet inorder to return a route error packet to it. If this host has an entry for the original sender in its route cache, itmay send the route error packet using that route. Otherwise, this host may reverse the route from the packetin error (the route by which the packet reached this host) or may use piggybacking as in the case of a routereply packet (Section 4.2). Another option in the case of returning a route error packet is for this host tosave the route error packet locally in a buffer, perform a route discovery for the original sender, and thensend the route error packet using that route when it receives the route reply for this route discovery. Thisoption cannot be used for returning a route reply packet, however, since then neither host would ever be able to complete a route discovery for the other, if either initially had a route cache entry for the other.Route maintenance can also be performed using end-to-end acknowledgements rather than the hop-byhopacknowledgements described above, if the particular wireless network interfaces or the environment inwhich they are used are such that wireless transmissions between two hosts do not work equally well in bothdirections. As long as some route exists by which the two end hosts can communicate (perhaps differentroutes in each direction), route maintenance is possible. In this case, existing transport or application level replies or acknowledgements from the original destination, or explicitly requested network level acknowledgements, may be used to indicate the status of this host's route to the other host. With hop-by-hopacknowledgements, the particular hop in error is indicated in the route error packet, but with end-to-endacknowledgements, the sender may only assume that the last hop of the route to this destination is in error.

4. SECURITY AND NODE COOPERATION

The wireless mobile ad hoc nature of MANETs brings new security challenges to the network design13. As the wireless medium is vulnerable to eavesdropping and ad hoc network functionality is established through node cooperation, mobile ad hoc networks are intrinsically exposed to numerous security attacks. During passive attacks, an attacker just listens to the channel in order to discover valuable information. This type of attack is usually impossible to detect, as it does not produce any new traffic in the network. On the other hand, during active attacks an attacker actively participates in disrupting normal operation of the network. This type of attackinvolves deletion, modification, replication, redirection and fabrication of protocol.control packets or data packets. Securing ad hoc networks against malicious attacks is difficult to achieve. Preventive mechanismsinclude among others authentication ofmessage sources, data integrity and protection of message sequencing, and aretypically based on key-based cryptography. Incorporating cryptographic mechanisms is challenging, as there is no centralised key distribution centre or trusted certification authority. These preventative mechanisms need to be sustained by detection techniques that can discover attempts to penetrate or attack the network. The previous problems were all related to malicious nodes that intentionally damage or compromise network functionality. However, selfish nodes, which use the network but do not cooperate to routing or packet forwarding for others in order not to spill battery life or network bandwidth, constitute an important problem as network functioning entirely relies on the cooperation between nodes and their contribution to basic network functions. To deal with these problems, the self-organising network concept must be based on an incentive for users to collaborate, thereby avoiding selfish behaviour. Existing solutions aim at detecting and isolating selfish nodes based on watchdog

mechanisms, which identify misbehaving nodes, and reputation systems,
which allow nodes to isolate selfish nodes. Another promising approach is the introduction of a billing system into the network based on economical models to enforce cooperation. Using virtual currencies or micro-payments, nodes pay for using other nodes' forwarding capabilities or services and are remunerated for making theirs available This approach certainly has potential in scenarios in which part of the ad hoc network and services is deployed by companies or service providers (for example, location- or context-aware services, sports stadium, taxi cab network, etc.). Also, when ad hoc networks are interconnected to fixed infrastructures by gateway nodes, which are billed by a telecom operator (for example, UMTS, hot-spot access, etc.), billing mechanisms are needed to remunerate these nodes for making these services available. Questions such as who is billing, to whom and for what, need to be answered and will lead to complex business models. We may conclude that in some ad hoc network scenarios, the network organization can completely or partially rely on a trust relationship between participating nodes (for example, PANs). In many others, security mechanisms, mechanisms to enforce cooperation between nodes or billing methods are needed and will certainly be an important subject of future research.

CONCLUSIONS

The rapid evolution in the field of mobile computing is driving a new alternative way for mobile communication, in which mobile devices form a self-creating, self-organising and self-administering wireless network, called a *mobile ad hoc network*. Its intrinsic flexibility, lack of infrastructure, ease of deployment, auto-configuration, low cost and potential applications make it an essential part of future pervasive computing environments. As a consequence, the seamless integration of mobile ad hoc networks with other wireless networks and fixed infrastructures will be an essential part of the evolution towards future fourthgeneration
communication networks. From a technological point of view, the realization of this vision still requires a large number of challenges to be solved related to devices, protocols, applications and services. The concise discussion in this paper shows that, despite the large efforts of the MANET research community and the rapid progress made during the last years, a lot of challenging technical issues remain unanswered. From an economical point of view, mobile ad hoc networks open up new business opportunities for telecom operators and
service providers. To this end, appropriate business scenarios, applications and economical models need to be identified, together with technological advances, making a transition of ad hoc networks.

REFERENCES:

[1] A Brief Overview of Ad Hoc Networks: Challenges and Directions: - Ram Ramanathan and Jason Redi, BBN Technologies. IEEE Communications Magazine, 50th Anniversary Commemorative Issue, May 2002

[2] Ad Hoc Networking:- Charles E. Perkins. Addison-Wesley, December 2000.

[3] Mobile Ad Hoc Networking: Imperatives and challenges: - Imrich Chlamtac, Marco Conti & Jennifer Liu. Ad Hoc Networks Publication, Volume 1, Issue 1, July 2003.

[4] On reducing the Broadcast Redundancy in Ad Hoc Wireless Networks: - Wei Lou & Jie Wu, IEEE transactions on mobile computing. April- June 2002.

[5] Secure message transmission in mobile ad hoc networks: Yih-Chun Hu; David B. Johnson; Adrian Perrig. Proceedings of the 2003 ACM workshop on Wireless security

[6] Security in Ad-hoc Networks: - Anne Vanhala, University of Helsinki, Department of Computing Science, Research seminar on Security in Distributed Systems.

[7] Securing Ad Hoc Networks: - Lidong Zhou & Zygmunt J. Haas, Cornell University, Ithaca, NY. IEEE Networks Special Issue on Network Security. November/December, 1999

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Analysis of Security issues and Bandwidth estimation techniques in MANET

[1]R.Rajeshkanna [2]Dr.A.Saradha

[1]Assistant Professor, Dr.N.G.P Arts & Science College, Coimbatore
[2]Head, Dept. Of CSE, Institute of Road and Transport Technology, Erode

ABSTRACT

Security is an important issue for ad hoc networks, especially for those security-sensitive applications. To secure an ad hoc network, we consider the following attributes: availability, confidentiality, integrity, authentication, and non-repudiation. Bandwidth estimation is an important issue in the Mobile Ad-hoc Network (MANET) because bandwidth estimation in MANET is difficult, because each host has imprecise knowledge of the network status and links change dynamically. Ad hoc networks present unique advanced challenges, including the design of protocols for mobility management, effective routing, data transport, security, power management, and quality-of-service (QoS) provisioning. Once these problems are solved, the practical use of MANETs will be realizable.

Keywords: Mobile Adhoc Netowrk, Security Issues, Routing Protocol Attacks, Bandwidth estimation

Introduction

Mobile ad hoc networks (MANETs) consist of a collection of wireless mobile nodes which dynamically exchange data among themselves without the reliance on a fixed base station or a wired backbone network. MANET nodes are typically distinguished by their limited power, processing, and memory resources as well as high degree of mobility. In such networks, the wireless mobile nodes may dynamically enter the network as well as leave the network. Due to the limited transmission range of wireless network nodes, multiple hops are usually needed for a node to exchange information with any other node in the network. Thus routing is a crucial issue to the design of a MANET.

The departure from traditional wired networks has brought about a shift in the primary concerns which need to be addressed when designing protocols of operation for wireless networks. While, like their wired counterparts, wireless networks are interested in high throughput and low-cost design, network priorities will dictate a tradeoff for an increase in power efficiency and bandwidth optimization. In particular, we consider adhoc networks. These peer-to-peer networks are composed of tens to hundreds of possibly homogeneous nodes with sensing and processing capabilities.

Bandwidth estimation is a basic function that is required to provide QoS in MANETs. It is a way to determine the data rate available on a network route. It is of interest to users wishing to optimize

end-to-end transport performance, overlay network routing, and peer-to-peer file distribution.

Techniques for accurate bandwidth estimation are also necessary for traffic engineering and capacity planning support. Having information existing can help to develop better methods for e.g. gateway selection, channel selection, routing, etc.

MANET nodes are typically differentiated by their limited power, processing, and memory resources as well as high degree of mobility. In MANETs, the wireless mobile nodes may dynamically enter in the network as well as leave the network. Because of the limited transmission range of wireless network nodes, multiple hops are generally required for a node to exchange information with any other node in the network.

Literature review

In an ad hoc network, a host's available bandwidth refers to amount of bandwidth available to the node to send packets to the network. Whole channel will not be used for packet transmission. Bandwidth estimation can be done using various methods; for example, bandwidth estimation is a cross-layer design of the routing and MAC layers and the available bandwidth is estimated in the MAC layer and is sent to the routing layer for admission control. Therefore, bandwidth estimation can be carried out in various network layers.

Present bandwidth estimation tools measure one or more of three related metrics: capacity, available bandwidth, and bulk transfer capacity. Currently available bandwidth estimation tools utilize a various strategies to measure these metrics. The issues of multipath routing in MANETs were specifically examined. They also discuss the application of multipath routing to support application constraints such as reliability, load-balancing, energy-conservation, and QoS. An improved mechanism was proposed to estimate the available bandwidth in IEEE 802.11-based ad hoc networks. In 802.11-based ad hoc networks, few works deal with solutions for bandwidth estimation. In a distributed ad hoc network, a host's available bandwidth cannot decided only by the raw channel bandwidth, but also by its neighbor's bandwidth usage and interference caused by other sources, each of which reduces a host's available bandwidth for transmitting data. Therefore, applications cannot properly optimize their coding rate without knowledge of the status of the entire network.

According to the simulations show that their QoS-aware routing protocol can improve packet delivery ratio greatly without impacting the overall end-to-end throughput, while also decreasing the packet delay and the energy consumption significantly.

The problem in available bandwidth estimation was rethink in IEEE 802.11 based ad hoc networks. According to them estimation accuracy is increased by improving the calculation accuracy of the probability for two adjacent nodes idle period to overlap.

SECURITY CHALLENGES IN AD HOC NETWORKS

Use of wireless links renders an Ad hoc network susceptible to link attacks ranging from passive eavesdropping to active impersonation, message replay and message distortion Eavesdropping might give an attacker access to secret information thus violating confidentiality. Active attacks could range from deleting messages, injecting erroneous messages; impersonate a node etc thus violating availability, integrity, authentication and non repudiation.

Nodes roaming freely in a hostile environment with relatively poor physical protection have non-

negligible probability of being compromised. Hence, we need to consider malicious attacks not only from outside but also from within the network from compromised nodes. Thus following are the ways by which security can be breached.

Vulnerability of Channels: As in any wireless network, messages can be eavesdropped and fake messages can be injected into the network without the difficulty of having physical access to network components.

Vulnerability of nodes: Since the network nodes usually do not reside in physically protected places, such as locked rooms, they can more easily be captured and fall under the control of an attacker.

Absence of Infrastructure: Ad hoc networks are supposed to operate independently of any fixed infrastructure. This makes the classical security solutions based on certification authorities and on-line servers inapplicable.

Dynamically Changing Topology: In mobile ad hoc networks, the permanent changes of topology require sophisticated routing protocols, the security of which is an additional challenge. A particular difficulty is that incorrect routing information can be generated by compromised nodes or as a result of some topology changes and it is hard to distinguish between the two cases.

Advantages of Ad Hoc Networks

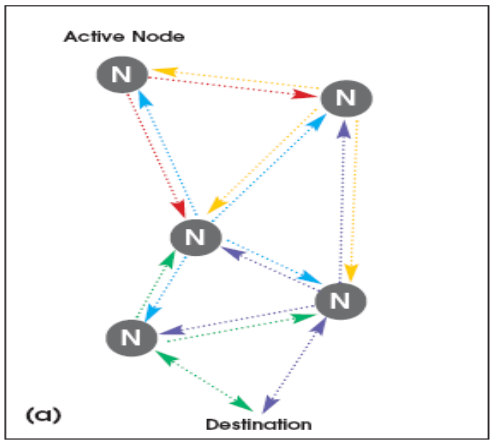
The principal advantages of an ad hoc network include the following:

- Independence from central network administration
- Self-configuring, nodes are also routers
- Self-healing through continuous re-configuration
- Scalable—accommodates the addition of more nodes
- Flexible—similar to being able to access the Internet from many different locations

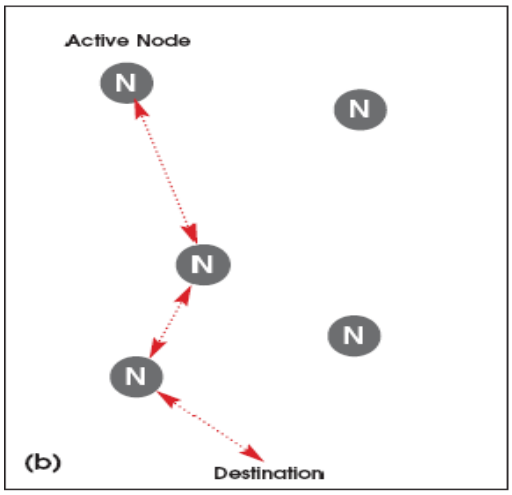
Key Applications

WLAN is the initial application that received a concerted development effort. Peer-to-peer networks of computer/ PDA users have become common. Commercial wireless Internet service providers (WISP) use repeater nodes to extend coverage to a large area, while user nodes can extend service in their locality. Control systems (e.g. environmental controls) and industrial process monitoring and control are becoming major applications for mesh networking. These environments are difficult to serve with dedicated wiring, being spread over a large area, often with difficult access.

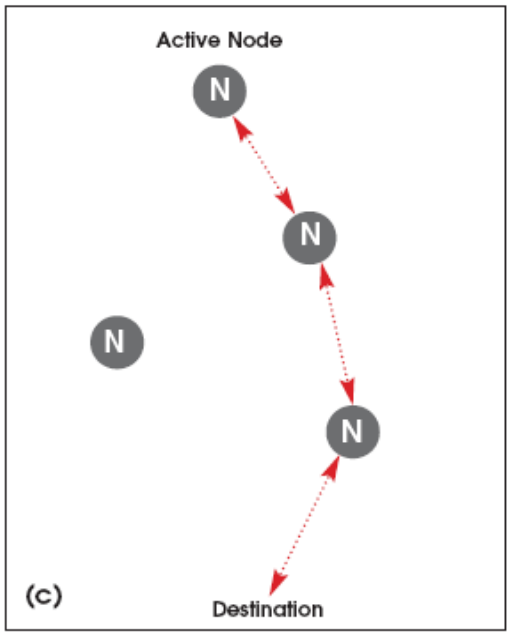
Sensor networks from small-scale (e.g. household security monitoring) to large scale (e.g. wildlife tracking) are also being developed with ad hoc networking as the operational structure. Developers of these and other applications have determined that ad hoc networks are the most efficient way to maintain system-wide communications. Hopefully, this brief tutorial will help you appreciate the value of this important networking technique for establishing and maintaining wireless communications.



a. Determination of available nodes



b. **Selection of the optimal routing**



c. **Reconfiguration when the network makeup changes**

Vulnerability in MANETs

Malicious and selfish nodes are the ones that fabricate attacks against physical, link, network, and application-layer functionality. Current routing protocols are exposed to two types of attacks:

- Active attacks
- Passive attacks

Classification of Security Attacks

Active attacks	Spoofing, Fabrication, Wormhole Attack, Modifi cation, Denial of Service
Passive Attacks	Eavesdropping, traffic analysis, monitoring

Active Attacks

Active attacks are the attacks that are performed by the malicious nodes that bear some energy cost in order to perform the attacks. Active attacks involve some modification of data stream or creation of false stream. These attacks can be classified into further following types.

Spoofing: Occurs when a malicious node misrepresents its identity in order to alter the vision of the network topology that a benign node can gather .

Fabrication: The notation “fabrication” is used when referring to attacks performed by generating false routing messages. Such kind of attacks can be difficult to identify as they come as valid routing constructs, especially in the case of fabricated routing error messages, which claim that a neighbor can no longer be contacted.

Wormhole Attack: An attacker records packets at one location in the network and tunnels them to another location. Routing can be disrupted when routing control messages are tunneled. This tunnel between two colluding attackers is referred as a wormhole. Wormhole attacks are severe threats to MANET routing protocols.

Modification: The attacker performs such attacks is targeted to integrity of data, by altering packet or modifying packets.

Denial of Service: This active attack aims at obstructing or limiting access to a certain resource. The resource can be a specific node or service or the whole network. The nature of ad-hoc networks, where several routes exist between nodes and routes are very dynamic gives ad hoc a built-in resistance to Denial of Service attacks, compared to fixed networks.

Passive Attacks

In passive attacks the attacker does not perturb the routing protocol, instead try to extract the valuable information like node hierarchy and network topology from it. Passive attack is in nature of eavesdropping on, or monitoring of, transmission. The goal of opponent is to obtained information that is being transmitted . Passive attacks are very difficult to detect because they do not involve any alteration of data.

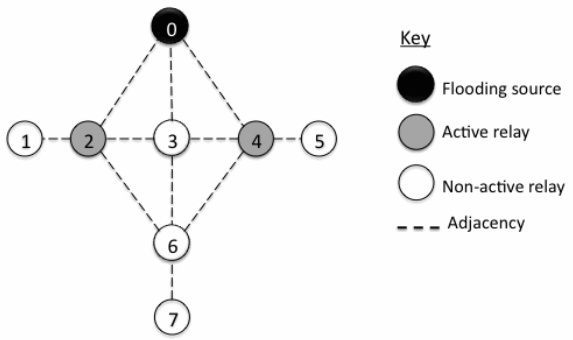
Other Advanced Attacks

We will now discuss several specific attacks that can affect the operation of a routing protocol in ad hoc network.

Blackhole attack: In a black hole attack a malicious node advertising itself as having a valid route to the destination. With this intension the attacker consume or intercept the packet without any forwarding . An attacker can completely modify the packet and generate fake information, this cause the network traffic diverted or dropped.

Byzantine attack: A compromised with set of intermediate, or intermediate nodes that working alone within the network carry out attacks such as creating routing loops, forwarding packets through non-optimal paths, or selectively dropping packets, which results in disruption or degradation of the routing services within the network .

Rushing attack: Two colluded attackers use the tunnel procedure to form a wormhole. If a fast transmission path (e.g. a dedicated channel shared by attackers) exists between the two ends of the wormhole, the tunneled packets can propagate faster than those through a normal multi-hop route. This forms the rushing attack . The rushing attack can act as an effective denial-of-service attack against all currently proposed on-demand MANET routing protocols, including protocols that were designed to be secure, such as ARAN and Ariadne .



Replay attack: An attacker that performs a replay attack are retransmitted the valid data repeatedly to

inject the network routing traffic that has been captured previously. This attack usually targets the freshness of routes, but can also be used to undermine poorly designed security solutions.

Location disclosure attack: An attacker discover the Location of a node or structure of entire networks and disclose the privacy requirement of network through the use of traffic analysis techniques, or with simpler probing and monitoring approaches. Adversaries try to figure out the identities of communication parties and analyze traffic to learn the network traffic pattern and track changes in the traffic pattern. The leakage of such information is devastating in security.

Bandwidth Estimation Methods

Estimating accurate available bandwidth allows a node to make optimal decision before transmitting a packet in networks. It is therefore clear that the available bandwidth estimation enhances the QoS in wired and wireless Networks. Measuring available bandwidth in ad hoc networks is challenging issue in MANET and calculating the residual bandwidth using the IEEE 802.11 MAC is still a challenging problem, because the bandwidth is shared among neighboring hosts, and an individual host has no knowledge about other neighboring hosts' traffic status. Two methods for estimating bandwidth are used below:

1. **"Listen" bandwidth estimation:** For hosts to listen to the channel and estimate the available bandwidth every second based on the ratio of free and busy times. The IEEE 802.11 MAC utilizes both a physical carrier sense and a virtual carrier sense [via the network allocation vector (NAV)], which can be used to find out the free and busy times. The MAC detects that the channel is free when the following three requirements are met:

• NAV's value is less than the current time;

• Receive state is idle;

• Send state is idle.

The MAC declares that the channel is busy when one of following occurs:

• NAV sets a new value;

• Receive state changes from idle to any other state;

• Send state changes from idle to any other state.

2. "Hello" bandwidth estimation: The sender's current bandwidth consumption as well as the sender's one-hop neighbors current bandwidth consumption is piggybacked onto the standard "Hello" message. Each host estimates its available bandwidth based on the information provided in the "Hello" messages and knowledge of the frequency reuse design.

The second neighboring host's information was proposed by using hop relay to propagate. **AODV** uses the "Hello" messages to update the neighbor caches. The "Hello" message used in AODV only keeps the address of the host who initiates this message. Modify the "Hello" message, including two fields. The first field includes host address, consumed bandwidth, timestamp, and the second field includes neighbor's addresses, consumed bandwidth, timestamp. Each host finds out its used bandwidth by monitoring the packets it supplies into the network. This value is recorded in a bandwidth-consumption register at the host and is updated periodically.

Conclusion

As wireless networks are becoming more

sophisticated and also offering more applications, number of security sensitive areas will appear depending on number of interfaces. There are a number of security efforts underway currently from both manufacturers and users, but still none of the individual security feature can encompass the meaning of security as a whole. MANET, attack types security criteria, which act as a guidance to the security-related research works in this area.

The proposed method, immediately releasing bandwidth when the route breaks as in the "Listen" method and replace this with an "update technique" used in the "Hello" method. The proposed method also does not cause an overhead which was in the "Hello" bandwidth Estimation method due to attaching neighbours bandwidth usage information.

References

[1] C.-K Toh, Ad Hoc Mobile Wireless Networks: Protocols and Systems, Prentice Hall, New Jersey, pp: 34-37, 2007.

[2] Larry Stotts, Scott Seidel, Tim Krout, and Paul Kolodzy, "MANET Gateways: Radio Interoperability via the Internet, Not the Radio", IEEE Communications Magazine, June 2008, Vol 46. No. 6, PP.51-60.

[3] Lt Col R J B Spencer, and Jim Ironside, "Network Centric Warfare Operation in an Expeditionary Context", Military Information & communications, Symposium of South Africa (MICSSA), July 2007.

[4] "Trust Management System for Network of Networks (NoN)",http://www.cse.dmu.ac.uk/STRL/research/nato/index.html, visited June 28, 2008.

[5] C. Siva Ram Murthy, and B.S. Manoj, Ad Hoc Wireless Networks: Architectures and Protocols, Prentice Hall communications engineering and emerging technologies series Upper Saddle River, New Jersey, 2004.

[6] Stephen Mueller, Rose P. Tsang, and Dipak Ghosal, "Multipath Routing in Mobile Ad Hoc Networks:Issues and Challenges" .

[7] Saleh Ali K.Al-Omari, Putra Sumari, "An Overview of Mobile Ad hoc Networks for the Existing Protocols And Applications", International Journal on Applications of Graph Theory in Wireless Ad hoc Networks and Sensor networks (Graph-Hoc), Vol.2, No.1, March 2010.

[8] Ali, Ahmed and Abdul Latiff, Liza and Fisal, Norsheila (2004) Indoor Location Tracking in Mobile Ad Hoc Network (MANET) using RSSI. In: RFM 2004, Oct 5-6, 2004, Subang, Selangor, Malaysia., "Indoor Location Tracking in Mobile Ad Hoc Network (MANET) using RSSI".

[9] R. S. Prasad, M. Murray, C. Dovrolis, K. Claffy, "Bamdwidth Estimation: Metrics, Measurement Techniques and Tools"

[10] Cheikh Sarr, Claude Chaudet, Guillaume Chelius, and Isabelle Gue´ rin Lassous, "Bandwidth Estimation for IEEE 802.11-Based ad Hoc Networks", IEEE Transactions on Mobile Computing, Vol. 7, No. 10, October 2008.

[11] Lei Chen, Student Member, IEEE, and Wendi B. Heinzelman, Member, IEEE, "QoS-Aware Routing Based on Bandwidth Estimation for Mobile Ad Hoc Networks", IEEE Journal on Selected Areas in Communications, Vol. 23, No. 3, March 2005.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

PERFORMANCE EVALUATION OF BANDWIDTH IN MANET

Mrs.B.Ramya

Assistant Professor, Department of Computer Applications

Dr. N.G.P Arts & Science College, Coimbatore

ABSTRACT

In wireless communication environments, backoff is traditionally based on the IEEE binary exponential backoff (BEB). In mobile ad hoc networks, channel bandwidth is a very limited resource. However no mechanism or method has been standardized to accurately evaluate the amount of resources remaining on a given channel. Such an evaluation would be a good asset for bandwidth-constrained applications. In multihop ad hoc networks, such evaluation becomes even more tedious. When the network is heavily loaded, the rate of collisions and hence the number of retransmissions increase. This leads to a notable decrease in network throughput and hence reduces quality of service because more and more bandwidth is wasted. MAC layer enables sharing the channel among all the nodes in the network effectively. It also enables perfect channel utilization. IEEE 802.11 DCF uses exponential backoff algorithm to reduce collision. In this, average backoff value is increased after every collision due the large size of contention window. The time consumed by backoff process is more. Hence, proportion of bandwidth is lost due to collision as well as the overhead is introduced by backoff algorithm.

Keywords: MANET, Mac Layer, Backoff, BEB.

1. INTRODUCTION

1.1 Binary Exponential Back off Algorithm

In the IEEE 802.11 standard MAC protocol, the BEB is used. This algorithm functions in the following way (Xu et al., 2002). When a node over the network has a packet to send, it first senses the channel, using a carrier sensing technique. If the channel is found to be idle and not being used by any other node, the node is granted access to start transmitting. Otherwise, the node waits for an interframe space, and the backoff mechanism is invoked. A random backoff time will be chosen in the range [0, CW-1]. A uniform random distribution is used here, where CW is the current contention window size. The following equation is used to calculate the backoff time (BO):

BO = (Rand () MOD CW)* aSlotTime. (1)

The backoff procedure is preformed then, by putting the node on a waiting period of length BO. Using carrier sense mechanism, the activity of the medium is sensed at every time slot. If the medium is found to be idle, then the backoff period is decremented by one time slot.

(BO) new = (BO) old – aSlotTime (2)

If the medium is determined to be busy during backoff, then the backoff timer is suspended, meaning that the backoff period is counted in terms of idle time slots. Whenever the medium is determined to be idle for longer than an

interframe space, backoff is resumed. When backoff is finished with a BO value of zero, a transfer should take place. If the node succeeded in sending a packet and receiving an acknowledgment for that packet, then the CW for this node is reset to the minimum, which is equal to 31 in the case of BEB. If the transfer fails, the node goes into another backoff period. When going for another backoff period again, the CW size is exponentially increased with a maximum of 1,023 (Zhai & Fang, 2003).

Disadvantages of BEB

One major disadvantage is the problem of fairness. BEB tends to prefer last contention winner and new contending nodes over other nodes when allocating channel access. This is done by choosing a random backoff value from a CW, which has a smaller size for new contending nodes and contention winners. This behavior causes what is known as the "channel capture effect" in the network. Another problem with BEB is stability. BEB has been designed to be stable for large numbers of nodes. Studies have shown that it is not.

step 0: set BO to initial value
step 1: while BO ≠ 0 do
For each time slot
If channel is idle then BO = BO -1
If channel idle for more than IdFs then
Send
Else
BO = log(BO) * BO
Go to step 1
step 2: Stop

1.2 The Logarithmic BackOff Algorithm

In IEEE 802.11 MAC protocol, the BEB algorithm exponentially increases the size of CW. BEB uses the following equation to increase the CW size.

BO = (Rand () MOD CW) * aSlotTime

(3)The logarithmic backoff algorithm has used the logarithm of the current BOas the incremental factor to calculate the next backoff. The following formula is used to give this result (Manaseer & Ould- Khaoua, 2005):

(BO) new = (log (BO) old) * (BO) old * aSlot-Time

(4)The used formula provides different outcome for backoff times, the behavior of the new formula can be seen in Figure 1. The further we go with backoff; the closer are the new values to the old values generated by the modified algorithm **(Manaseer & Ould-Khaoua, 2005).**

The main idea behind choosing such an equation for calculating BO is that instead of going on a backoff period for X time slots , the node goes into two consecutive backoff periods, say i1 and i2, where i1 + i2 ≈ X, when the node is on a backoff period for a consecutive number of times. This allows the node a chance to access the channel and transmit in a way**.**

Figure 1. CW increase in logarithmic algorithm

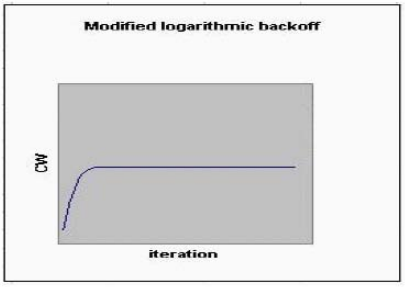


Figure 2:. The logarithmic backoff algorithm.

2. IMPLEMENTATION AND RESULTS

The logarithmic algorithm has been evaluated and compared against the standard BEB using the network simulator NS 2.33 **(Fall & Varadhaa, 2002)**.

Network Simulator NS-2

In this project NS-2 is chosen as the simulation environment because it is the most widely used network simulator. The Network

Simulator (NS), is a discrete event simulator for networks, is a simulated program developed by VINT (Virtual InterNEtwork Testbed) project group (A Collaboration among USC/ISI, Xerox PARC LBNL and UCB). It supports simulation of TCP and UDP, some of MAC layer protocols, various routing and multicast protocols over both wired and wireless network etc. The TCL script is given as the input to the Network Simulator. The TCL interpreter generates the results using the input code and the NS Library. The Simulation results can beviewed using NAM (Network Animator)

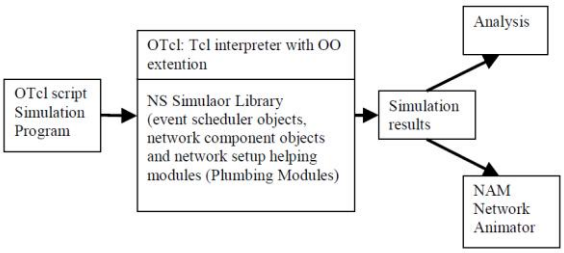


Figure 3: Structure for Network Simulator

NS-2 has several key benefits which suit for doing the simulation:

It provides extensive support for simulating TCP/IP, Routing and multicast protocols over wired and wireless networks It provides a lot of standard modules to be used in mobile networks.

It is an object oriented design which provides a lot of documents.

It use TCL to specify the components and OTcl to glue them together.

2.1 Implementation

The original standard MAC protocol has been modified to implement the logarithmic algorithm. Modifications have mainly targeted the mac802_11.h and mac802_11.cc files Several topologies and mobility scenarios have been created to test the algorithm as intensively as possible. Firstly, we have varied the total number of nodes in the network. Changing the number of nodes is used to predict the performance of our algorithm for all size networks. Simulations have been carried out for networks having a total number of nodes varying between 20 and 100 mobile nodes.Other simulation parameters are an area of 600m×600m, simulation time of 100 seconds, and nodes transmission range of 250 m, and the traffic generated is constant bit rate (CBR) traffic at a rate of 1 packet per second.

Table 1: General Parameters for Simulation Environment

Parameter	Values
Nodes	50
Simulation Time	100s
Hello intervals	1s
Grid Size	600m x 600m
Traffic type	CBR
Medium Capacity	2.5 Mb/s

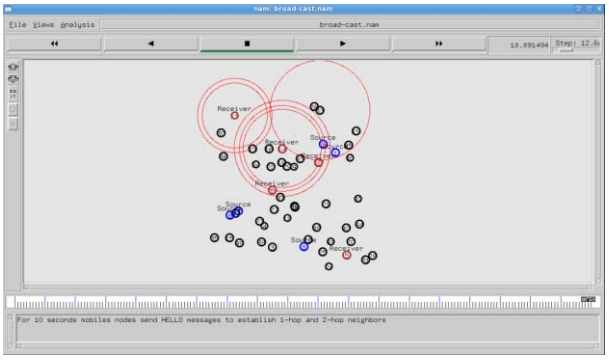


Figure 4: Network Animator for BEB

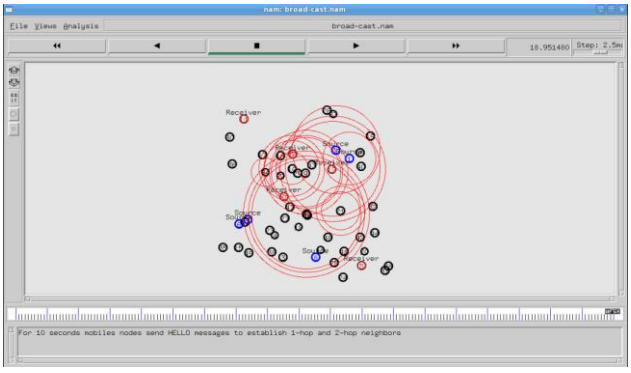


Figure 5: Network Animator for LOG

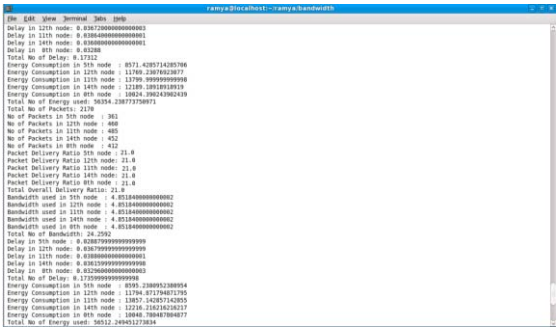


Figure 6: Simulation result of BEB

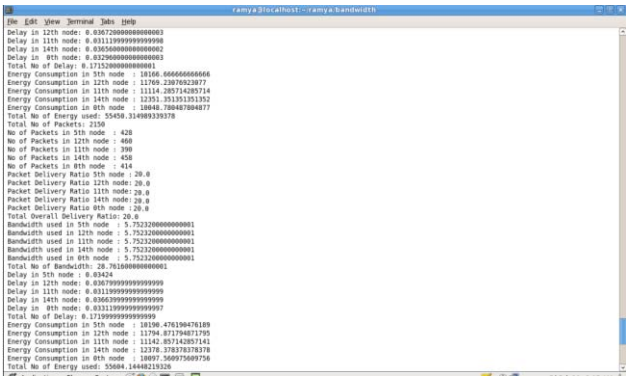


Figure 7: Simulation result of LOG

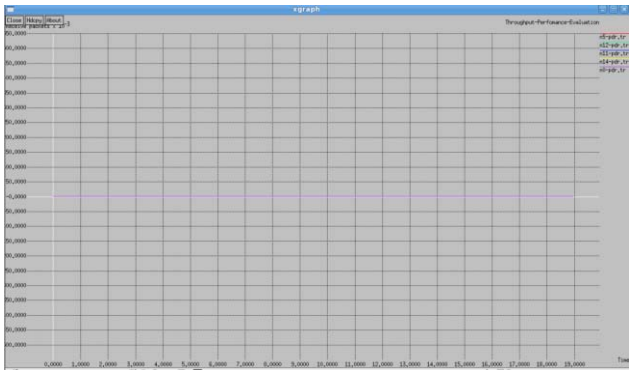


Figure 8: Throughput of BEB

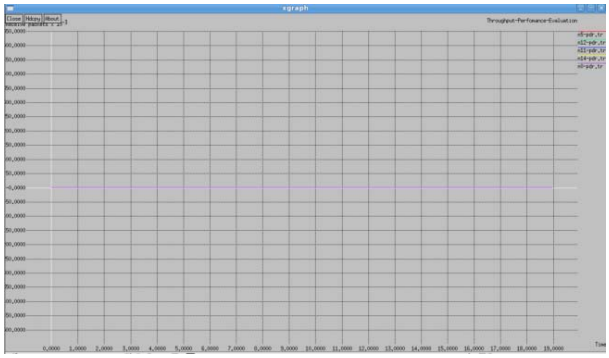


Figure 9: Throughput of LOG

Nodes	BEB Throughput	LOG Throughput
Node 5	21	20
Node 12	21	20
Node 11	21	20
Node 14	21	20
Node 0	21	20

Table2: Showing the throughput values of BEB and Log

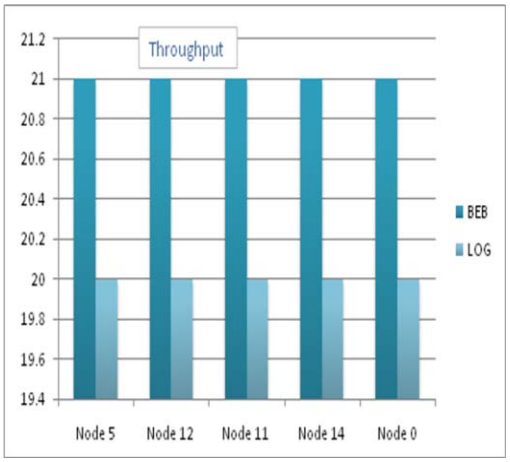


Figure 10: Comparison chart of Throughput for BEB and LOG

The Figure 21 displays the results of running the Logarithmic algorithm against the standard IEEE 802.11 Binary Exponential Backoff algorithm. The figure shows that the throughput is higher for the modified algorithm. A network with a larger number of nodes has a better throughput, than one with a small number of nodes. The reason for this is that for a larger number of nodes, contention is much higher, so it is more probable for a node to backoff for more consecutive periods. This leads to a more significant effect of the behavior of the logarithmic algorithm, and the backoff values start to be closer.

The BEB uses a uniform random number distribution generator. The random distribution used covers the effect of CW incremental behavior. The reason for excluding the random number distribution in our algorithm is that a formal distribution only benefits from the window size increment by 50% of the generated random values, which is useless in the case of the modified algorithm. The graph in Figure 4 shows a

comparison of the same modified algorithm with another version of itself, where a random number distribution is used in the same way it is used by the standard binary exponential backoff algorithm.While many researches try to analyze and understand the incremental behavior of a backoff algorithm, the effects of all the changes made are reduced by the random number distribution

sed.

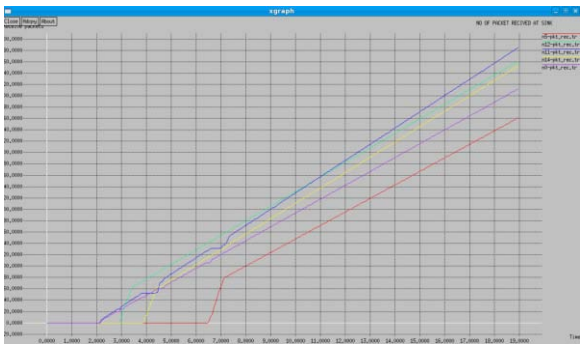


Figure 11: Forwarded Data Packet for BEB

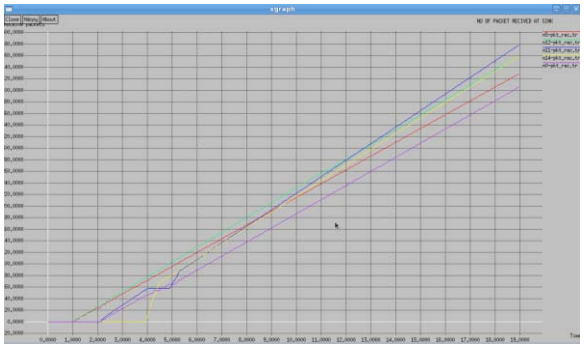


Figure 12: Forwarded Data Packet for LOG

Nodes	BEB Packets Forwarded	LOGPackets Forwarded
Node 5	361	428
Node 12	450	460
Node 11	360	390
Node 14	452	458
Node 0	412	414

Table 3: Showing the packets forwarded valuesofBEB

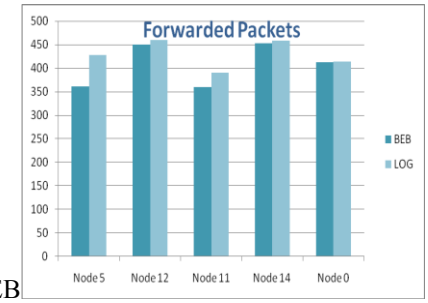


Figure 13: Comparison chart of Forwarded Packets for BEB and LOG

Figure 24 shows the packet forwarding count as an indicator of node activityand channel usage. Once again, the logarithmic algorithm achieves a higher number of forwarded packets almost over all nodes used for this experiment.

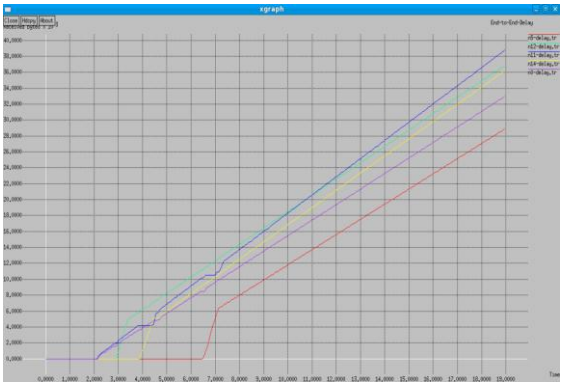


Figure 14: End to End Delay for BEB

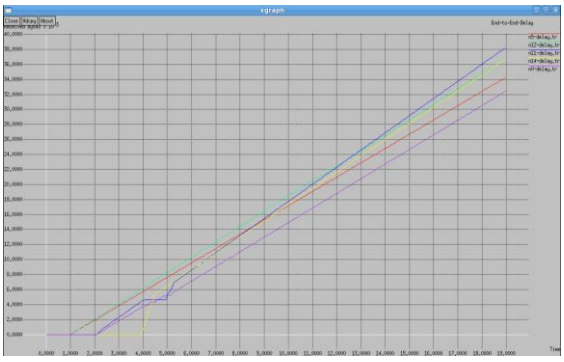
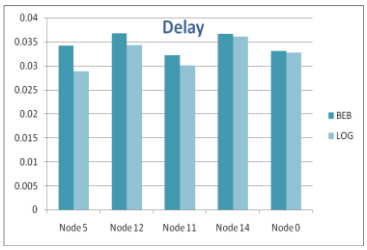


Figure 15: End to End Delay for LOG

Nodes	BEB Delay	LOG Delay
Node 5	0.03424	0.02887
Node 12	0.03679	0.03432
Node 11	0.03219	0.03012
Node 14	0.03663	0.03615
Node 0	0.03311	0.03273

Table 5: Showing the delay values of BEBand

Figure 16: Comparison chart of Delay for BEB and LOG

In the simulation, we studied the effect of mobility speed on average nod 

e-to-node delay. As expected, the BEB exponential increment causes longer idle time, increasing the average delay over the network. As seen in Figure 27, LOG has reduced average delays. Increasing network size leads to a larger number of CW size increments, since there are more colliding nodes.

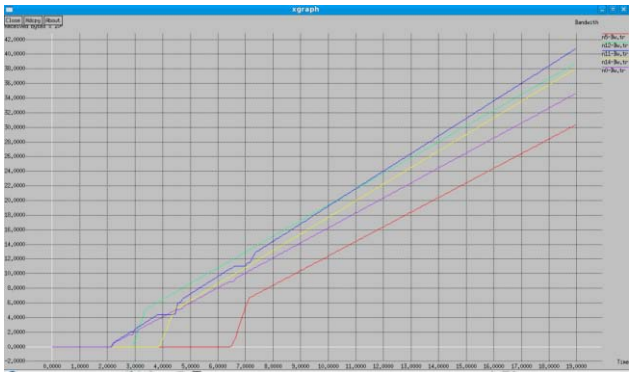


Figure 17: Bandwidth for BEB

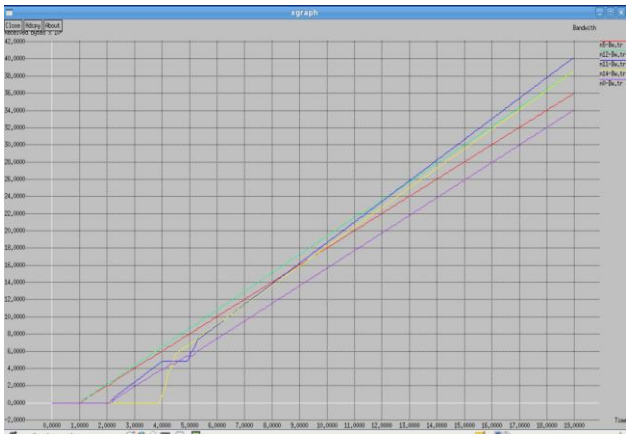


Figure 18: Bandwidth for LOG

Nodes	BEB Bandwidth	LOG Bandwidth
Node 5	8595.23	10190.47
Node 12	11794.87	11986.87
Node 11	13857.14	14530.85
Node 14	12216.21	12378.37
Node 0	10048.78	10097.56

Table 6: Showing the Bandwidth values of BEB and Log

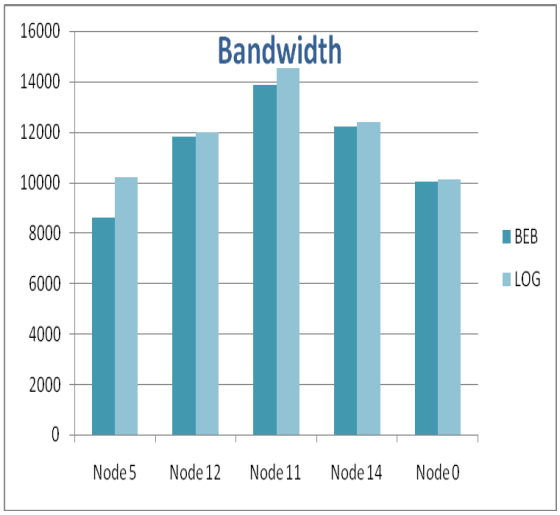


Figure 19: Comparison chart of Bandwidth for BEB and LOG

The simulation results of estimating the bandwidth used or energy consumed is high in the logarithmic backoff than that of BEB algorithm as shown in the Figure 30. This shows the efficient usage of the bandwidth.

5. Conclusions and Future Work

The BEB is used by the IEEE 802.11MAC protocol. BEB uses uniform random distribution to choose the backoff value. In this paper, we have studied a modified logarithmic backoff algorithm, which uses logarithmic increments, instead of the exponential extension of window size to eliminate the effect of random number distribution. Results from simulations have demonstrated that the modified algorithm increased the total throughput of the MANETs, especially when the system size is large. Moreover, the tested algorithm also has shown stable network

throughput and increased packet forwarding over the network. Added to that, the logarithmic algorithm also has decreased the average node-to-node delay, especially.

One drawback of the studied algorithm is a slight increment in the number of dropped packets during simulation time when the network size is large. But this drawback is caused by other factors as well. Studying these factors is left for future work by introducing other backoff algorithms for better results.

3.References

1. Andreas Johnsson, Bob Melander, and Mats Björkman. Bandwidth Measurement in Wireless Network. Technical report,

2. Bao, L., & Garcia-Luna-Aceces, J. J. (2001). A new approach to channel access scheduling for ad hoc networks. ACM MOBICOM, 210-221.

3. Bononi, L., Budriesi, L., Blasi, D., Cacace, V., Casone, L., & Rotolo, S. (2004). A differentiated distributed coordination function MAC protocol for cluster-based wireless ad hoc networks. In Proceedings of the 1st ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, And Ubiquitous Networks (pp. 77-86).

4. S.Chakrabbarti, A.Mishra, "QoS Issues in Ad Hoc Wireless Networks", IEEE Communication Magazine, Feb2001.

5. Charles E. Perkins and Pravin Bhagwat. Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers. SIGCOMM Comput. Commun. Rev., 24(4):234–244, 1994.

6. T. Clausen and P. Jacquet. Request for comments: 3626, optimized link state routing protocol (OLSR). http://www.ietf.org/rfc/rfc3626.txt, October 2003.

7. R. de Renesse, M. Ghassemian, V. Friderikos, A.H Aghvami. Adaptive Admission Control for Ad Hoc and Sensor Networks Providing Quality of Service,. Technical report, King College London, May 2005

8. C. Dovrolis, P. Ramanathan, and D. Moore. What do Packet Dispersion Techniques Measure? In Proceedings of IEEE INFOCOM, pages 905–914, Apr.2001 showed that this is not the case.

9. Fang, Z., Bensaou, B., & Wang, Y. (2002).Performance evaluation of a fair backoff algorithm for IEEE 802.11 DFWMAC. International Symposium on Mobile Ad Hoc Networking & Computing, (pp. 48–57).

10. Frank Y. Li, Mariann Haugea, Andreas Hafslund, Oivind Kure and Pal Spilling. Estimating Residual Bandwidth in 802.11-based Ad Hoc Networks: An empirical Approach. In Proceedings of The Seventh International Symposium on Wireless Personal Multimedia Communications (WPMC 2004), Abano Terme, Italy, September 2004

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A review study on Artificial Immune System and its Applications

[1]B.Ramya [2]Dr.K.Arthi [3]Somasundaram

[1] Assistant Professor, Department of Computer Applications , Dr.N.G.P Arts and Science College
[2]Director [3] Assistant Professor , RVS College of Arts and Science,Coimbatore

ABSTRACT

A miracle of our body, the enemy cells like bacteria or viruses are prevented by Immune System from entering the body. This biological Immune System (IS) has the capacity of detecting or recognizing self/non-self antigen in the body. Artificial Immune System (AIS) is one of the recent biologically inspired approaches to emerge from computer science, process nonlinear classification properties along with biological properties such as self/non-self identification, Clonal Selection and Negative Selection. During the last decade, the field of Artificial Immune System (AIS) is progressing slowly and steadily as a branch of computational Intelligence (CI).There has been increasing interest in the development of computational models inspired by several Biologically Inspired Systems. In particular, some are building models mimicking the mechanisms in the biological immune system (BIS) to better understand its natural processes and simulate its dynamical behaviour in the presence of antigens/pathogens. The AIS models are using simplified models of various immunological processes and functionalities for designing artifacts–computational algorithms, techniques

Keywords – artificial immune system, artificial immune networks, clonal selection, negative selection

I. INTRODUCTION

AIS have been defined by [de Castro & Timmis 2002a] as:
"adaptive systems, inspired by theoretical immunology and observed immune functions, principle and models, which are applied to problem solving"

Various aspects of biology have always been the inspiration in developing computational models and problem solving methods. The immune system is one such system that has recently drawn significant attention; and as a result, the Artificial Immune System (AIS), has emerged Figure (1). The powerful information processing capabilities of the immune system, such as feature extraction, pattern recognition, learning, memory, and its distributive nature provide rich metaphors for its artificial counterpart. The immune system, however, is a system with high complexity and is under active research (from the biological point of view), likewise the current AIS works adopted only a few immune mechanisms. Specifically, three immunological principles are primarily used in a piecemeal in AIS methods. These include the immune network theory, the mechanisms of negative selection, and the clonal selection principles. This paper tries to summarize the works on existing models and development of new ones.[1]

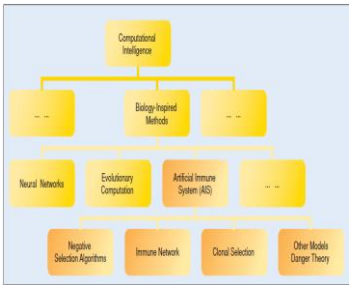


Figure (1) : Artificial Immune System (AIS) as a branch of Computational Intelligence (CI). From [2]

II. BIOLOGICAL BASE COMPUTATIONAL SYSTEMS: THEORETICAL BACKGROUND

A. The Immune System

The vertebrate immune system is composed of diverse sets of cells and molecules that work together with other systems (like neural and endocrine) for maintaining homeostatic state. The primary function of the immune system is to protect human bodies from infectious agents (such as viruses, bacteria and other parasites) commonly known as *pathogens*. Immune response is incited by the recognition of an associated molecule called

antigen. Immune system usually works according to two mechanisms namely: Innate and Adaptive Immunity. Innate immunity is directed against general pathogens that enter the body while adaptive or acquired immunity allows launching an attack against any invader that innate system cannot remove. For more information about the immune systems, the reader can refer to [3], [4], [5].

B. Negative Selection Mechanism

Negative selection is a mechanism to protect body against self-reactive lymphocytes It utilizes the immune system's ability to detect unknown antigens while not reacting to the self cells viewed in Figure [2]. During the generation of T-cells, receptors are made through a pseudo-random genetic rearrangement process. Then, they undergo a censoring process in the thymus, called the negative selection. In this process, T-cells that react against self-proteins are destroyed and only those that do not bind to self-proteins are allowed to leave the thymus. These matured T-cells then circulate throughout the body performing immunological functions and protecting the body against foreign antigens [7].

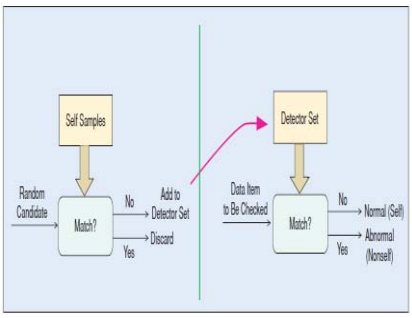


Figure [2] : The basic concept of the Negative Selection (NS) Algorithm.

C. The Clonal Selection Principle

Clonal selection theory was proposed by Burnet (1959). The theory is used to explain basic response of adaptive immune system to antigenic stimulus. It establishes the idea that only those cells capable of recognizing an antigen will proliferate while other cells are selected against. Clonal selection operates on both B and T cells. B cells, when their antibodies bind with an antigen, are activated and differentiated into *plasma* or memory cells. Prior to this process, clones of B cells are produced and undergo somatic hyper mutation. As a result, diversity is introduced into the B cell population. Plasma cells produce antigen-specific antibodies that are work against antigen. Memory cells remain with the host and promote a rapid secondary response[4].

The main features of the Clonal Selection Theory are that: ·

- The new cells are copies of their parents (clone) subjected to a mutation mechanismwith high rates (somatic hypermutation);
- Elimination of newly differentiated lymphocytes carrying self-reactive receptors;
- Proliferation and differentiation on contact of mature cells with antigens.

III. ARTIFICIAL IMMUNE SYSTEMS: LITERATURE REVIEW

A. Clonal Selection Based Algorithms

The Clonal Selection principle is the whole process of antigen recognition, cell proliferation and differentiation into memory cell [8]. Several artificial immune algorithms have been developed imitating the clonal selection theory.

Castro and Zuben (2002) [3] proposed a clonal selection algorithm named CLONALG for learning and optimization, CLONALG generates a population of N antibodies, each specifying a random solution for the optimization process. During each iteration, some of the best existing antibodies are selected, cloned and mutated in order to construct a new candidate population. New antibodies are then evaluated and certain percentage of the best antibodies is added to the original population. Finally a percentage of worst antibodies of previous generation are replaced with new randomly create ones.

In [9] an Immunity Clonal Strategy algorithm (ICS) which includes Immunity Monoclonal Strategy Algorithms (IMSA) and Immunity Polyclonal Strategy Algorithm (IPSA) is introduced. ICS is used to solve multi-objective optimization task. Zuo and Li (2003) [10] proposed a Chaos Artificial Immune Algorithm (CAIF) for function optimization problem. It uses chaotic variable to perform local search and explore the solution space.

Garrett (2004) [11] introduced an Adaptive Clonal Selection (ACS) algorithm as a modification of CLONALG. It suggests some modifications to the CLONALG based on an analysis of the operators for selecting the amount of mutation and number of clones to overcome the drawbacks of CLONALG such as the several parameters used and the binary representation. An Adaptive Immune Clonal Strategy Algorithm (AICSA) proposed for solving numerical optimization problems in [12]. It assigns dynamically the immune memory unit and antibody population according to the Ab-Ab and Ab-Ag affinities. It also integrates the local search with the global search.

An adaptive clonal algorithm proposed in [16] for optimal phasor measurement unit (PMU) placement. It adjusts the number of the cycle supplement population and the probabilities of hypermutation and recombination operators of the CLONALG algorithm. These modifications can enhance the optimization process and help to avoid the locally optimal traps. Cutello et al. (2005) [15] introduced an improved version of OPT-IA called opt-IMMALG. The main modifications in this algorithm are the replacement of the binary string representation by a real-coded one and the introduction of a new inversely proportional hyper mutation operator.

B. Negative Selection Based Algorithms

The Negative Selection is one of the mechanisms of the natural immune system that has inspired the developments of most of the existing Artificial Immune systems. In the T-cell maturation process of the immune system, if a T-cell in thymus recognizes any self cell, it is eliminated before deploying for immune functionality. Similarly, the negative selection algorithm generates detector set by eliminating any detector candidate that match elements from a group of self samples.

Negative selection based algorithms have been used in different applications areas, such as anomaly detection. Forrest (1994) [26], proposed a negative selection algorithm. The main idea of his algorithm is to generate a set of detectors by first randomly making candidates and then discarding those that recognize training self-data, and then these detectors can later be used to detect anomaly.

C. Artificial Immune Network Models

Based on the immune network theory proposed by Jerne (1974) [6], which was presented in the previous section, many researchers have developed models that use ideas and concept from the immune network theory to solve problems in different application areas. The pioneer work of Ishiguro et al. (1994) [33] and Hunt and Cooke (1996) [34] have motivated many researchers and hence several models have been developed in the literature. Following the work done in Dasgupta et al (2003) [1], we will summarize some of the existing Immune network models in this section with focus on the last five years.

An Artificial Immune System for E-mail Classification (AISEC) presented by Secker et al. (2003) [42]. It is capable of continuous learning for the purpose of two class classification and used for the task of electronic mail sorting. Alonso et al. (2004) [43] proposed based on the aiNet model an approach to model an agent that plays the Iterated Prisoner's Dilemma (IDP). The agent structure consists of two immune networks: recognition AIN and a decision AIN. The main improvement to the aiNet is introduced in the mechanism the network uses to add B-cell to the memory.

In Lv (2007) [51] a Chaos Immune Network (CIN) algorithm for multimodal function optimization is discussed. The main features of this algorithm are the use of chaos variable to simulate proliferation mode of immune cells to enhance searching accuracy, the stepping criteria was improved and some relevant measures have been added to avoid pre-maturation. Huang and Jiao (2007) [52] presented an Artificial Immune Kernel Clustering Network (IKCN) for unsupervised image segmentation. It combines the artificial immune network and the support vector domain description. In this model the Image features sets will be divided into subsets by the antibodies and then each subset mapped into a hypersphere in a high dimensional feature space by a Mercer Kernel. Finally, a minimal spanning tree is used to automatically determine the final number of clusters without a predefined number of clustering. Some other proposed immune network algorithms can be found in Li et al. (2008) [53].

IV. CONCLUSION

An overview of the Artificial Immune Systems field including a theoretical background on the main ideas and concepts of AIS and the recent advances in the literature have been presented in this survey. This has provided a motivation to continue exploring the AIS field and contribute to the development of the new AIS models and techniques.

.

Although AIS models have achieved great successes in various application domains, there are still some theoretical issues that need to be further explored such as the development of unified frameworks, convergence, and scalability. The developments of the artificial immune systems would benefit not only from the inspiration of biological immune principles and mechanisms, but also hybridization with other soft computing paradigms, such as neural networks, fuzzy logic, and genetic algorithms. They could also be further studied and applied to more challenging application areas and to solve complex real world problems.

Table 2 show a chronological list of some AIS models and techniques developed in the literature since Dasgupta's et al. (2003) work. A brief description for each model or technique, the aspect of the biological immune systems modelled, the type of representation used and the application area to which AIS has been applied are included in the table.

TABLE 2: A TIME-LINE OF AIS WORKS

Reference	Model or technique description	Aspects of the BIS modelled	Type of representation used	Applications
(Neal, 2003)	Meta-stable memory immune system for multivariate data analysis.	Immune Networks	Real-valued	Data analysis
(Rouchen et al., 2003)	An Immunity Clonal Strategy Algorithm (ICS) to solve multi-objective optimization task.	Clonal Selection	Real-valued vectors	Optimization
(Zuo and Li, 2003)	A Chaos Artificial Immune Algorithm (CAIF) by integrating of chaotic search and CLONALG	Clonal Selection	Real-valued vectors	Optimization
(Nasraoui et al., 2003)	Techno – streams model for detecting an unknown number of evolving clusters in a noisy data stream	Immune Networks	Real-valued	Clustering
(Secker et al., 2003)	An artificial immune system for E-mail classification (AISEC)	Immune Networks	Two-part words vector	Classification
(Garrett, 2004)	An Adaptive Clonal Selection (ACS) algorithm that suggests some modifications to the CLONALG	Clonal Selection	Real-valued vectors	Optimization
(Gonzalez and Canady, 2004)	A self adaptive negative selection algorithm for anomaly detection.	Negative Selection	Binary strings, real-valued	Anomaly Detection
(Yu and Hou, 2004)	An improved Clonal selection algorithm based in CLONALG	Clonal selection Ag-Ab binding	Binary Strings	Machine Learning
(Liu et al., 2004)	An Adaptive Immune Clonal Strategy Algorithm (AICSA)	Ag-Ab binding Clonal Selection	Real-valued vectors	Numerical Optimization problems
(Bentley and Timmis, 2004)	A Fractal immune networks model combining the ideas of fractal proteins with immune networks.	Immune Networks	Real-valued	Classification, Clustering
(Luh and Lin, 2004)	A Reactive Immune Network (RIN) for mobile robot learning navigation strategies within unknown environments	Immune Networks	Real-valued	Robots
(Campels et al., 2005)	A Real-coded Clonal selection Algorithm (RCSA) that enable the treatment of real valued variables for optimization problems.	Clonal Selection	Real-valued vectors	Electromagnetic design optimization
(Xian et al., 2005)	A novel unsupervised Fuzzy K-Means (FKM) clustering anomaly detection algorithm based on clonal selection algorithm.	Clonal Selection	Numeric characteristic variables	Computer Security
(Cutello et al., 2005)	Immunological algorithm for continuous global optimization problems name OPI-IA	Clonal Selection	Binary String	Optimization
(Cutello et al., 2006)	An improved version of OPT-IA called Opt – IMMALG	Clonal Selection	Real-code	Optimization
(Qiao and Jianping, 2006)	An Immune based Network Intrusion Detection System (AINIDS)	Immune Networks	Rules	Computer Security
(Bian and Qiu, 2006)	An adaptive Clonal algorithm that suggests some modifications to the CLONALG	Clonal selection, receptor editing	Binary strings	Optimization
(Karakasis et al., 2006)	A hybrid model which combines clonal selection principles and gene expression programming	Clonal selection	Symbol Strings	Data Mining
(Tian et al., 2006)	A modified algorithm of aiNet to solve function optimization problems	Immune Networks	Real-valued vector	Optimization
(Hao and Cai-Xin, 2007)	Artificial immune network classification algorithm (AINC) for fault diagnosis of power transformer.	Immune Networks	Real-valued	Classification
(Zhang and Yi, 2007)	A Tree structured artificial immune network (TSAIN) model for data clustering and classification.	Immune Networks, Clonal Section	Real-valued	Classification, Clustering
(Fu et al., 2007)	A hybrid artificial immune network that uses the swarm learning	Immune Networks	Real-valued	Optimization
(Lv, 2007)	A chaos immune network algorithm combines chaos idea with immune network to improve its ability of searching peaks.	Immune Networks,	Real-valued	Optimization
(Zeng et al., 2007)	A feedback negative selection algorithm (FNSA) for anomaly detection.	Negative Selection	Real-valued	Anomaly Detection

(Huang and Jiao, 2007)	An Artificial Immune kernel clustering network (IKCN) for unsupervised image segmentation.	Immune Networks	Real-valued, Image features sets	Clus terin g
(Gan et al., 2007)	A technique that combines gene expression programming with clonal selection algorithm for system modelling & knowledge discovery.	Clonal selection	Symbol Strings, Binary String	Syst em Mod ellin g
(Graaff and Engelbre cht, 2007)		Immune Networks	Real-valued	Clus terin g

REFERENCES

[1] D. Dasgupta, Z. Ji & D. Dasgupta, Z. Ji "Artificial Immune System (AIS) Research in the Last Five Yea'rs" IEEE Evolutionary computation , 2003. CEC'03 The 2003 Congress on (Volume 1) Pages 123-130

[2] Dipankar Dasgupta, University of Memphis, USA "Advances in Artificial Immune Systems" published in IEEE Computational Intelligence Magazine | November 2006

[3] Castro L. de and Zuben F., 1999. Artificial Immune Systems: Part I – Basic Theory and Applications. TR – DCA 01/99.

[4] Castro L. de and Timmis J., 2003. Artificial Immune Systems as a Novel Soft Computing Paradigm. Soft Computing Journal, vol. 7, Issue 7.

[5] Timmis J., Knight T., Castro L. de and Hart E., 2004. An Overview of Artificial Immune Systems, In R Paton, H Bolouri, M Holcombe, J H Parish, and R Tateson, editors, "Computation in Cells and Tissues: Perspectives and Tools for Thought", Natural Computation Series, pages 51-86. Springer.

[6] (Jerne1974), N. K. Jerne, Towards a network theory of the immune system. Ann. Immunol, (Inst. Pasteur), 125C, pp373-389, 1974.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

OPTIMIZING SCORE AND DECISION FUSION IN PALM AND FINGER PRINTS USING PARETO OPTIMIZATION TECHNIQUES

M.Ramya M.Phil Scholar

Department of Computer Science KG college of Arts and Science Coimbatore

ABSTRACT

A multimodal biometric system integrates information from multiple biometric sources to compensate for the limitations in performance of each individual biometric system. Different approaches have been proposed in the literature for developing unimodal and multimodal biometric systems. Multi-biometric systems are developed by fusing different biometric features pertaining to various biometric modalities at different levels. Many researchers have shown that the multimodal systems outperform the unimodal systems, giving better discrimination of genuine from imposters. In existing system they we present multimodal systems at feature level and score level fusions using our already reported unimodal palm print and fingerprint identifiers. The unimodal finger- and palm print identification systems utilize directional energies of texture as features, extracted using contourlet transform. To improve the optimization in result it can propose the pareto-optimal search method in order to handle the top-k query in the high-dimension record set with that the nearest neighbor search is found and it is used to make dominant relationship between them while fusion. After that the scores of palm print and fingerprint images are multiplied together to produce a new set of values consisting of combined values of both the systems. Then the total number of generated score of test image corresponding to trained database obtained from that decision making will be done.

Keywords Multimodal, matching, scores, summation, fusion

1. INTRODUCTION

The multi-biometrics system can be developed by utilizing much variety of approaches. The multi-sensor systems are made by the combination of different types of sensors under a single attribute. The system providing combination of algorithm is processed through single biometric modality using multiple algorithms. The system providing multi instance combines multiple instances of the same attribute. Some system uses multiple samples of same bio-metric modality using a single sensor. The multimodular systems are developed by integrating the information of different biometric behaviour of the individual to establish unique identity. The database is trained at first. At score level fusion it matches the scores generated through system by comparing input test images. A normalized euclidean distance classifier is setup which is used to generate the matching scores. Different features of palmprint and fingerprint are compared with their respective databases using this classifier. These scores contain less amount of information as compare to feature values. Before integrating the scores together, these scores should be normalized to a common scale. In both palmprint and fingerprint systems normalized energy values are used to generate the scores. These generated scores are already on a common scale and thus eradicate the need of using any score normalization technique. The palm and finger scores are joint using two rules: summation Rule and invention Rule

2. FUSION METHODOLOGY

2.1 Feature Extraction Level

This extraction level comes under pre-classification fusion. The data obtained from each sensor is used to obtain a feature value. The success of a feature extraction algorithm for classification problems is determined by the probability of misclassification. The feature obtained from one biometric attribute is not dependent of those extracted from the other sources. Different values are obtained from different sources and finally these features are combined to form one pattern. It is further used for classification.

2.2 Matching Scores Level

This level comes under post classification fusion. The approach shifts the scores of the spoofing attacks in an attempt to create a margin between them and the score distribution of the valid users. The scores obtained from different classifier provide a matching score indicating the closeness of the feature value with the pattern value. Based on this classification is performed. These scores can be united to declare the authenticity of the claimed identity.

2.3 Decision Level

This level also comes under post classification fusion. In this level an additional check is performed using the verification system and then a final decision is taken. The multiple biometric data is capture using sensor and the resulting feature values is classified individually into two module–acceptance or rejection. In multimodal biometric systems score level fusion is commonly preferred because matching scores contain enough information to make genuine and impostor case distinguishable. These values are relatively easy to acquire.The matching scores for a pre-specified number of users can be generated in a given number of biometric systems even these is insufficient knowledge of the underlying feature extraction and matching algorithms seen in each system. Therefore, integrating information using score level fusion obtained from individual modalities seems both viable and practical. Even though the scores obtained from a biometric system can be either similar scores or distinct scores; it is converted into a same nature.

Let X denotes some of the raw matching scores from a definite matcher, and let xϵX. The normalized score of x is then denoted by x‘. These normalization schemes can be used to both summation rule-based fusion and SVM- based fusion for improving accuracy.

2.4 Min–Max Normalization

The normalization maps the raw matching scores to interval and retains the original distribution of matching scores except for a scaling factor. Given that max(X) and min(X) are the maximum and minimum values of the raw matching scores.

3. FINGERPRINT AND PALMPRINT RECOGNITION

The fingerprint and palm print seen in human are unique and permanent throughout a person’s life. In this there is a fusion of minutia score matching method for fingerprint and minutia alignment matching algorithm for palmprints. A person’s fingerprint is normally comprised of ridges and valleys. The dark area of the fingerprint is known as ridges. The white area that exists between the ridges is known as valleys. There are many types of classifications seen in different patterns that can arise in the ridges. These points are also known as the minutiae of the fingerprint. In current fingerprint recognition technologies the most commonly used minutiae are ridge endings and bifurcations because they can be easily detected only by only looking at points that surround them.

Most of the modern fingerprint matching technologies use minutiae matching. If we find enough minutiae in one image and the same type of minutiae in another image then the images are taken from the same fingerprint of one person. The design of minutiae decides the recognition of images. The minutiae of same fingerprints are usually matched together by their space comparative to other minutiae around it. If the multiple points of fingerprints seen in one image have similar distances between them compared to the multiple points in another image then the images are said to match up with each other.

It is the proposal of this paper to add the constraint that the regions and ends between the minutiae should be approximately the same as it should be. The minutia code and alignment-based minutia matching algorithm is used to match two palmprints. A match score is estimated and calculated using the local ridge direction and frequency seen in the palmprints. The characteristic information around each minutia is calculated using the fixed length minutiae descriptor.

3.1 Palm Feature Extractor

In this module the image attainment setup is provided with two flat plate images. The camera and the light sources needed are fixed on the upper plate. The bottom plate is used to place the hand for image attainment with fixed pegs. The mismatch due to scale variance is minimized by keeping the distance between these two plates constant. The distance between the plates is kept based on the formula. The palmprint image is binarized using Hysteresis thresholding through isolating the foreground of palmprint from the background. The binarized palmprint is complemented and distance transform from the light source is calculated. The distance between that pixel and the nearest nonzero pixel assigns a number for the distance transform based on each pixel. The maximum distance acheived from the distance make over is estimated as the centre of palmprint. An effort is made to acquire standard palmprint images during image acquisition stage. A rotational alignment is incorporated in our proposed approach to cater any involuntary small rotations. The longest line in a palm passes through the middle finger. The rotation is measured based on this line. The second-order moment helps analyzing the elongation or eccentricity of any binary shape. An Eigen value is determined to find the idiosyncrasy of the shape by analyzing the ratio of the Eigen values.

3.2 Finger Print Feature Extractor

In this module the input image is pre-processed using histogram equalization, adaptive thresholding, the Fourier renovate and adaptive

binarization. To find the region of interest (ROI) from the input image, a core point is calculated and used as the reference point in the images. The core point is the point in the finger where the inner most ridges have the maximum curvature.

The region of interest (ROI) of a defined pixels size found around the core point is extracted from input image. With the help of Directional Filter Banks a 2-D spectrum is split into fine slices. A typical verification system on its own is trained to discriminate only between two classes: valid users as a positive and impostors as a negative class. In the work three approaches are investigated: score-level fusion, decision-level and feature-level fusion of verification system.

4. FEATURE LEVEL FUSION

The feature value present already is matched with the stored multimodal database. The success of a feature extraction algorithm for classification problems is determined by the probability of misclassification. The matching in the database consists of Euclidian classifier. The threshold value decides whether the result obtained is a genuine or impostor. Based on unimodal identifier the features obtained are matched with database using a Euclidian classifier. The decision is taken based on the selected threshold in the decision module.In this feature level fusion of fingerprint and palmprint, the feature values obtained from palmprint and fingerprints are joined together to make combined feature value.

Let $P = pp1, pp2 \ldots ppm$ and $F = fp1, fp2 \ldots fpn$ symbolize feature values presentating the information obtained from palmprint and fingerprint. The purpose is to combine these two feature value sets after normalization in a way to yield a joint feature value (JFV). JFV is obtained by integrating P and F feature value obtained from the sets. The problem of compatibility of feature sets is surmounting because values both in palm and fingerprint identifiers consist of normalized energy values. Finally the need for normalizing feature value sets is removed. JFVs are generated and stored in order to make multimodal database which is consequently used for identification and verification purpose thereafter.

5. SCORE LEVEL FUSION

The fusion seen at score level needs the matching scores derived by comparing input test image with trained database. The feature value sets of palmprint and fingerprint are compared with the concerned databases using euclidean distance classifier which is used to generate the matching scores. These scores include less amount of information as compare to feature values. As it contain only less information scores should be normalized to a common scale value before fusing or combining scores together.The normalized energy values are used in both palmprint and fingerprint systems to generate the scores. Thus the generated scores are already on a common scale and hence no score normalization technique is needed. The approach shifts the scores of the spoofing attacks in an attempt to create a margin between them and the score distribution of the valid users.

It is calculated by using entropy of an image which is as follows.

-sum (p.*log2 (p))

The palm and finger scores are combined using two rules: summation rule and invention rule.

5.1. Summation Rule

In the summation rule, the score value obtained from the palmprint and fingerprint input images are added together to defer a new set of values to get more data. Thus, the new set of values contains more amount of information as compared to the normal previous systems. Thus it gives more information about a person identity. Finally, the decision done on the value is based on the input claim that is established on the basis of preset threshold by the classifier. Suppose P = pp1, pp2, . . . ppm and F = fp1, fp2 . . . *fpn* give the scores of palmprints and fingerprints images correspondingly. According to the summation rule, the generated score value $sskk$ is obtained. It is calculated as $sskk = ppkk + ffkk$. Here, 'k' = the total number of score obtained from the test image related to trained database already present. sk denotes the combined generated score which is used mainly in decision making.

5.2. Invention Rule

The scores obtained from the palmprint and fingerprint images are multiplied with each other to produce a new set of values consisting of combined values generated from both the sets. Suppose $P = pp1, pp2 \ldots ppm$ and $F = fp1, fp2 \ldots fpn$ give the scores of palm and finger images, correspondingly. Thus in the product rule the combined score value ck is obtained. It is calculated as: c$k = pk + fk$ Here, 'k' = the total number of score obtained from the test image related to trained database already present. ck denotes the combined generated score which is used mainly in decision making.

6. FUSION BASED ON PARETO-OPTIMAL SEARCH METHOD

In this module scores should be normalized to a common scale value before fusing scores together. The normalized energy values are used in both palmprint and fingerprint systems to generate the scores. Thus the generated scores are

already on a common scale and hence no score normalization technique is needed. The palm and finger scores are combined using two rules: summation Rule and invention Rule. Based on these a decision making process is considered and executed. This is done by using the pareto-optimal search method. This method is the combination of features present in feature level and score level in which a dominating feature is found and estimated. From these finding the nearest neighboring feature and the dominating feature for scheduled values are noted and decision making process is evaluated.

7. PERFORMANCE EVALUATION

The images were collected using Digital Persona scanner, with the help of developed platform. A database consisting of palm and finger images of 40 individuals can be taken. 10 prints are collected from single person with 8 records per biometric modality. The multimodal database consists of 10×40 = 400 records. It consists of 200 palmprint and 200 fingerprint records. The database should be developed in two sessions with an average interval of two months to focus on performance of developed multimodal system. The database is trained before in both the phase for both palm and fingerprints. In our experiments, the developed database is divided into two sets: training and validation sets of 200 images. The validation data set is then used to estimate the performance of trained system.

7.1 Precision

In the field of information retrieval, precision is the fraction of retrieved image that are relevant to the search:

$$\text{Precision} = \frac{|\{relevant\ images\} \cap \{retrieved\ images\}|}{|\{retrieved\ images\}|}$$

The precision takes all retrieved image into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. The measure is called precision at n or P@n.

For example for a text search on a set of image precision is the number of correct results divided by the number of all returned results. The precision is also used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system.

7.2 Accuracy

The accuracy is the proportion of the total number of predictions that were correct. It is determined using the equation:

a is the number of correct predictions that an instance is negative,

b is the number of incorrect predictions that an instance is positive,

c is the number of incorrect of predictions that an instance negative, and

d is the number of correct predictions that an instance is positive.

$$\text{Accuracy} = a+d/a+b+c+d$$

7.3 Recall

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Recall} = \frac{|\{relevant\ images\} \cap \{retrieved\ images\}|}{|\{relevant\ images\}|}$$

For example for image search on a set of images recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant images also, for example by computing the precision.

The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP = d/c+d$$

The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$FP = b/a+b$$

7.4 F-Measure Comparision

It is used to compare how similar two clustering results are:

To begin lets have two clustering results [14]:

C = the correct vector of bit masks = {c1, c2, …..cn}

K= the vector of bit mask results of some algorithm = {k1, k2, k3….km}

Then create a “matching matrix”, M = [a_{ij}] . The matching matrix is just the number of cells

that from result C are in clustering i, and from result K are in cluster j. So i go from 1→n, and j goes from 1→m. N is then the total number of cells. So imagine that the correct result C divides the data set into 3 clusters. Then use a clustering algorithm that produces result K and divides the data into 2 clusters. The matching matrix M would be a 3 x 2 matrix beginning in the top left hand box, all the cells that were in correct cluster 1 and the clustering algorithm put into cluster 1.

The F measure is then a measure of the algorithms, precision and recall.

F measure= (2 * precision * recall) / (precision + recall), where:

Precision (P) = cells correctly put into a cluster / total cells put into the cluster

Recall (R) = cells correctly put into a cluster / All the cells that should have been in the cluster.

7.5 ROC Comparison

A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate). It shows the tradeoff between sensitivity and specificity. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

8. CONCLUSION

Under the framework, a hybrid fusion method is proposed, which integrate the score-level fusion and the decision-level fusion. It takes the advantage of both the palm and fingerprint. By defining the pareto- optimal features in a feature mining phase we can guarantee that the fusion method achieved and implemented here is the best for the personal identification. It uses a combination of depth first traversals of the network to find the pareto-optimal feature efficiently. The results show that the proposed system has the potential to identify all pareto-optimal features with a small percentage of feature evaluations.

9. REFERENCES

[1] R.V. Tappeta and J.E. Renaud, Interactive Multiobjective Optimization Procedure, American Institute of Aeronautics and Astronautics Journal, 1999.

[2] R.V. Tappeta, J.E. Renaud, A. Messac and G.J. Sundararaj, Interactive Physical Programming: Tradeoff Analysis and Decision making in Multidisciplinary Optimization, American Institute of Aeronautics and Astronautics Journal, 2000.

[3] R.V. Tappeta, J.E. Renaud and J.F. Rodr´ıguez, An Interactive Multiobjective Optimization Design Strategy for Decision Based Multidisciplinary Design, Engineering Optimization, 2002.

[4] K. Klamroth and K. Miettinen, Integrating Approximation and Interactive Decision Making in Multicriteria Optimization, Operations Research, 2008.

[5] Y.Y. Haimes and V. Chankong, Kuhn-Tucker Multipliers as Trade-Offs in Multiobjective Decision-Making Analysis, Automatica, 1979.

[6] H. Yano and M. Sakawa, Trade-Off Rates in the Weighted Tchebycheff Norm Method, Large Scale Systems, 1987.

[7] H. Nakayama, Trade-Off Analysis Using Parametric Optimization Techniques, European Journal of Operational Research, 1992.

[8] I. Kaliszewski and W. Michalowski, Generation of Outcomes with Selectively Bounded Trade-Offs, Foundations of Computing and Decision Sciences, 1995.

[9] I. Kaliszewski and W. Michalowski, Efficient Solutions and Bounds on Tradeoffs, Journal of Optimization Theory and Applications, 1997.

[10] H. Kuk, T. Tanino and M. Tanaka, Trade-Off Analysis for Value Optimization problems via Scalarization, Journal of Information & Optimization Sciences, 1997.

[11] M.I. Henig and J.T. Buchanan, Tradeoff Directions inMultiobjective Optimization Problems, Mathematical Programming, 1997.

[12] J.-B. Yang and D. Li, Normal Value Identification and Interactive Tradeoff Analysis Using Minimax Formulation in Multiobjective Optimization, IEEE Transactions on Systems, Man, and Cy- [1] R.V. Tappeta and J.E. Renaud, Interactive Multiobjective Optimization Procedure, American Institute of Aeronautics and Astronautics Journal, 1999.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

RESEARCH ON DATA MINING USING NEURAL NETWORK TRAINING

Mrs.Renuka Devi.S.V **M.Phil Scholar**

Department of Computer Science, Dr.SNS RCAS, Coimbatore.

ABSTRACT

Data mining refers to extracting or mining knowledge from large amounts of data. If we see the data that have been stored in our data warehouses and databases like a mountain, the gems are buried within the mountain. The companies are asking is how to extract this value. The answer is Data mining. Data mining helps end users to extract useful business information from large databases. There are many technologies available to data mining practitioners, including Artificial Neural Networks, Regression, and Decision Trees. Many practitioners are wary of Neural Networks due to their black box nature, even though they have proven themselves in many situations. Data mining is a multidisciplinary field bridging many technical areas such as databases technology, statistics, artificial intelligence, machine learning, pattern recognition and data visualization methods. Data mining is an essential step in the process of Knowledge Discovery in Databases (KDD). This paper is an overview of artificial neural networks and questions their preferred by tool of data mining practitioners.

Index Terms: Artificial Neural Network (ANN), Neural network Topology, back propagation, KDD.

1. Introduction

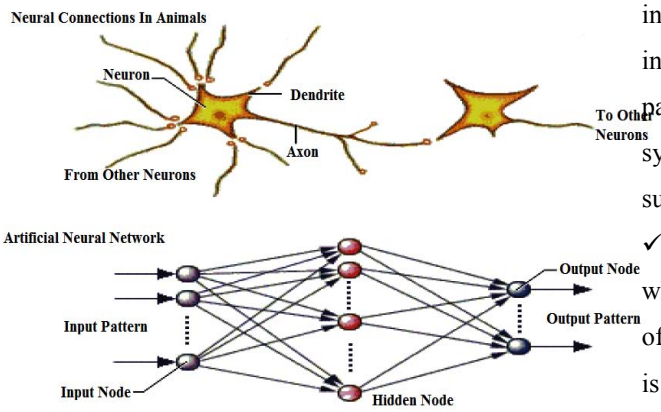
Data mining is the term used to describe the process of extracting value from a database. A data-warehouse is a location where information is stored. The type of data stored depends largely on the type of industry and the company. Many companies store every piece of data they have collected, while others are more ruthless in what they deem to be important. Consider the following example of a financial institution failing to utilize their data-warehouse. Another example of where this institution has failed to utilize its data-warehouse is in cross-selling insurance products (e.g. home, life and motor vehicle insurance).

By using transaction information they may have the ability to determine if a customer is making payments to another insurance broker. This would enable the institution to select prospects for their insurance products. These are simple examples of what could be achieved using data mining. Four things are required to data-mine effectively: high quality data, the right data, an adequate sample size and the right tool. There are many tools available to a data mining practitioner. These include decision trees, various types of regression and neural networks. Income is a very important socio-economic indicator. If a bank knows a person's income, they can offer a higher credit card limit or determine if they are likely to want information on a home loan or managed investments. Even though this financial institution had the ability to determine a customer's income in two ways, from their credit card application, or through regular direct deposits into their bank account, they did not extract and utilize this information.

2. Artificial Neural Networks

An artificial neural network (ANN), often just called a neural network(NN), is a mathematical model or

computational model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.



2.1 Neural Network Topologies

Feed-forward neural network

The feed-forward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network. The data processing can extend over multiple (layers of) units, but no feedback connections are present, that is, connections extending from outputs of units to inputs of units in the same layer or previous layers.

Recurrent network

Recurrent neural network that contain feedback connections. Contrary to feed-forward networks, recurrent neural networks (RNs) are models with bi-directional data flow. While a feed-forward network propagates data linearly from input to output, RNs also propagate data from later processing stages to earlier stages.

2.2 Training Of Artificial Neural Networks:

A neural network has to be configured such that the application of a set of inputs produces (either 'direct' or via a relaxation process) the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a priori knowledge. Another way is to 'train' the neural network by feeding it teaching patterns and letting it change its weights according to some learning rule. We can categorize the learning situations as follows:

✓ **Supervised learning** or Associative learning in which the network is trained by providing it with input and matching output patterns. These input-output pairs can be provided by an external teacher, or by the system which contains the neural network (self-supervised).

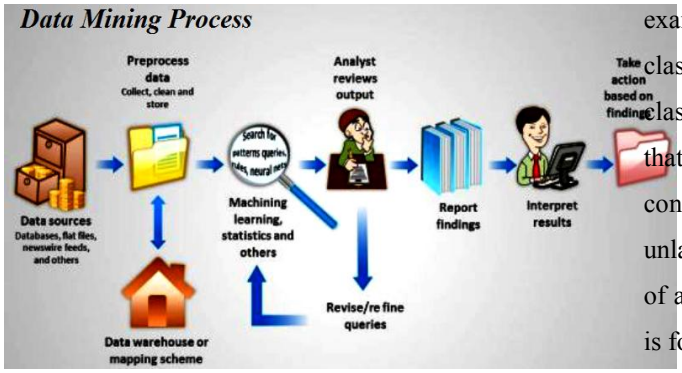
✓ **Unsupervised learning** or Self-organization in which an (output) unit is trained to respond to clusters of pattern within the input. In this paradigm the system is supposed to discover statistically salient features of the input population. Unlike the supervised learning paradigm, there is no a priori set of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli.

✓ **Reinforcement Learning** This type of learning may be considered as an intermediate form of the above two types of learning. Here the learning machine does some action on the environment and gets a feedback response from the environment.

3. Neural Networks In Data Mining

In more practical terms neural networks are non-linear statistical data modelling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. The difference between these data warehouses and ordinary databases is that there is actual manipulation and cross-fertilization of the data helping users makes more informed decisions. Neural networks essentially comprise three pieces: the architecture or model; the learning algorithm; and the

activation functions.



Neural networks are programmed or trained to store, recognize, and associatively retrieve patterns or database entries; to solve combinatorial optimization problems; to filter noise from measurement data; to control ill-defined problems; in summary, to estimate sampled functions when we do not know the form of the functions. It is precisely these two abilities (pattern recognition and function estimation) which make artificial neural networks (ANN) so prevalent a utility in data mining. As data sets grow to massive sizes, the need for automated processing becomes clear. With their model-free estimators and their dual nature, neural networks serve data mining in a myriad of ways. Data mining is the business of answering questions that you've not asked yet.

Data mining reaches deep into databases. Data mining tasks can be classified into two categories: Descriptive and predictive data mining. Descriptive data mining provides information to understand what is happening inside the data without a predetermined idea. Predictive data mining allows the user to submit records with unknown field values, and the system will guess the unknown values based on previous patterns discovered form the database.

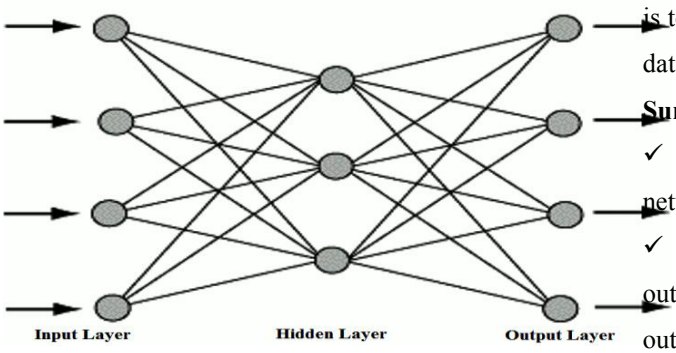
Data mining models can be categorized according to the tasks they perform: Classification and Prediction, Clustering, Association Rules. Classification and prediction is a predictive model, but clustering and association rules are descriptive models.

The most common action in data mining is classification. It recognizes patterns that describe the group to which an item belongs. It does this by examining existing items that already have been classified and inferring a set of rules. Similar to classification is clustering. The major difference being that no groups have been predefined. Prediction is the construction and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of a given object is likely to have. The next application is forecasting. This is different from predictions because it estimates the future value of continuous variables based on patterns within the data. Neural networks, depending on the architecture, provide associations, classifications, clusters, prediction and forecasting to the data mining industry. Financial forecasting is of considerable practical interest.

Due to neural networks can mine valuable information from a mass of history information and be efficiently used in financial areas, so the applications of neural networks to financial forecasting have been very popular over the last few years. Some researches show that neural networks performed better than conventional statistical approaches in financial forecasting and are an excellent data mining tool. In data warehouses, neural networks are just one of the tools used in data mining. ANNs are used to find patterns in the data and to infer rules from them. Neural networks are useful in providing.

3.1. Feed-forward Neural Network

One of the simplest feed-forward neural networks (FFNN), such as in Figure, consists of three layers: an input layer, hidden layer and output layer. In each layer there are one or more processing elements (PEs). PEs is meant to simulate the neurons in the brain and this is why they are often referred to as neurons or nodes. A PE receives inputs from either the outside world.



The simplified process for training a FFNN is as follows:

✓ Input data is presented to the network and propagated through the network until it reaches the output layer. This forward process produces a predicted output.

✓ The predicted output is subtracted from the actual output and an error value for the networks is calculated. The neural network then uses supervised learning, which in most cases is back propagation, to train the network. Back propagation is a learning algorithm for adjusting the weights. It starts with the weights between the output layer PE's and the last hidden layer PE's and works backwards through the network.

✓ Once back propagation has finished, the forward process starts again, and this cycle is continued until the error between predicted and actual outputs is minimized.

3.2. The Back Propagation Algorithm:

Back propagation or propagation of error, is a common method of teaching artificial neural network how to perform a given task. The back propagation algorithm is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers, and send their signals forward, and then the errors are propagated backwards. The back propagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the back propagation algorithm is to reduce this error, until the ANN *learns* the training data.

Summary Of The Technique:

✓ Present a training sample to the neural network.

✓ Compare the network's output to the desired output from that sample. Calculate the error in each output neuron.

✓ For each neuron, calculate what the output should have been, and a *scaling factor*, how much lower or higher the output must be adjusted to match the desired output. This is the local error. Adjust the weights of each neuron to lower the local error.

Actual Algorithm:

1. Initialize the weights in the network (often randomly)
2. Repeat

*For each example e in the training set do

1. = neural-net-output (network, e) ; forward pass
2. T = teacher output for e
3. Calculate error (T - O) at the output units
4. Compute delta_wi for all weights from hidden layer to output layer ; backward pass
5. Compute delta_wi for all weights from input layer to hidden layer ; backward pass continued
6. Update the weights in the network

* End

3. Until all examples classified correctly or stopping criterion satisfied return (network)

4. Review Of Literature Reporting neural Network Application

There are numerous examples of commercial applications for neural networks. These include;

✓ *Select Appropriate Paradigm:* Decide on network architecture according to general problem area (e.g., Classification, filtering, pattern recognition, optimization, data compression, prediction), Decide on transfer function, Decide on learning method, Select network size. Eg. How many inputs and output neurons? How many hidden layers and how many

neurons per layer? Decide on nature of input/output. Decide on type of training used.

✓ *Select Input Data and Facts:* Decide the problem domain, the training set should contain a good representation of the entire universe of domain .select input sources and optimal size of training set.

✓ *Data Preparation:* The next step is to think about different ways to represent the information. Data can be non-distributed or distributed. Using a NON Distributed date set, each neuron represents 100% of an item. The data set must be non-overlapping and complete. In this case, the network can represent only a limited number of unique patterns. Using a Distributed data set, the qualities that define a unique pattern are spread out over more than one neuron.

5. Advantage Of Neural Networks

- High Accuracy: Neural networks are able to approximate complex non-linear mappings
- Noise Tolerance: Neural networks are very flexible with respect to incomplete, missing and noisy data.
- Independence from prior assumptions: Neural networks do not make a priori assumptions about the distribution of the data, or the form of interactions between factors
- Ease of maintenance: Neural networks can be updated with fresh data, making them useful for dynamic environments.
- Neural networks can be implemented in parallel hardware
- When an element of the neural network fails, it can continue without any problem by their parallel nature.
- Assign blame for the local error to neurons at the previous level, giving greater responsibility to neurons connected by stronger weights.
- Repeat the steps above on the neurons at the previous level, using each one's blame as its error fraud detection, telecommunications, medicine, marketing, bankruptcy, prediction, insurance and the list goes on. The following are examples of where neural networks have been used.

Accounting

✓ Identifying tax fraud

✓ Enhancing auditing by finding irregularities

Finance

✓ Signature and bank note verification Risk Management

✓ Foreign exchange rate forecasting Bankruptcy prediction

✓ Customer credit scoring

✓ Credit card approval and fraud detection

✓ Forecasting economic turning points

✓ Bond rating and trading

✓ Loan approvals

✓ Economic and financial forecasting

Marketing

✓ Classification of consumer spending pattern

✓ New product analysis

✓ Identification of customer characteristics

✓ Sale forecasts

✓ Human resources

✓ Predicting employee's performance and behaviour

✓ Determining personnel resource requirements

6. Design Problems

There are no general methods to determine the optimal number of neurones necessary for solving any problem. It is difficult to select a training data set which fully describes the problem to be solved. Designing Neural Networks using Genetic Algorithms Neuro-Fuzzy Systems

7. Conclusion

There is rarely one right tool to use in data mining. It is a question as to what is available and what gives the

best results. Many articles, in addition to those mentioned in this paper, consider neural networks to be a promising data mining tool. Artificial Neural Networks offer qualitative methods for business and economic systems that traditional quantitative tools in statistics and econometrics cannot quantify due to the complexity in translating the systems into precise mathematical functions. Hence, the use of neural networks in data mining is a promising field of research especially given the ready availability of large mass of data sets and the reported ability of neural networks to detect and assimilate relationships between a large numbers of variables.

In most cases neural networks perform as well as better than the traditional statistical techniques to which they are compared. Resistance to using these black boxes is gradually diminishing as more researchers use them, in particular those with statistical backgrounds. Thus, neural networks are becoming very popular with data mining practitioners, particularly in medical research, finance and marketing. This is because they have proven their predictive power through comparison with other statistical techniques using real data sets. Due to design problems neural systems need further research before they are widely accepted in industry. As software companies develop more sophisticated models with user-friendly interfaces the attraction to neural networks will continue to grow.

References

[1] Agrawal, R., Imielinski, T., Swami, A., "Database Mining: A Performance Perspective", *IEEE Transactions on Knowledge and Data Engineering*, pp. 914-925, December 1993

[2] Berry, J. A., Lindoff, G., Data Mining Techniques, Wiley Computer Publishing, 1997 (ISBN 0-471-17980-9).

[3] Berson, "Data Warehousing, Data-Mining & OLAP", TMH

[4] Bhavani,Thura-is-ingham, "Data-mining Technologies,Techniques tools & Trends", CRC Press

[5] Bradley, I., Introduction to Neural Networks, Multinet Systems Pty Ltd 1997.

[6] Fayyad, Usama, Ramakrishna " Evolving Data mining into solutions for Insights", communications of the ACM 45, no. 8

[7] Fausett, Laurene (1994), Fundamentals of Neural Networks: Architectures, Algorithms and Applications, Prentice-Hall, New

Jersey, USA.

[8] Haykin, S., *Neural Networks*, Prentice Hall International Inc., 1999

[9] Khajanchi, Amit, Artificial Neural Networks: The next intelligence.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

APPLICATIONS & TRENDS IN DATA MINING

P.Rizwan Ahmed
Head & Asst. Professor
Department of Computer Applications and PG Department of Information Technology
Mazharul Uloom College, Ambur – 635802, Tamil Nadu

ABSTRACT

The advent of computing technology has significantly influenced our lives and two major impacts of this effect are Business data Processing and Scientific Computing. During the early years of the development of computer techniques for business, computer professionals were concerned with designing files to store the data so that information could be efficiently retrieved. There were restrictions on storage size for storing data and on the speed of accessing the data. Needless to say, the activity was restricted to a very few, highly qualified professionals. Then came an era when Database Management System simplified the task. The responsibility of intricate tasks, such as declarative aspects of the programs was passed on to the database administrator and the user could pose his query in simpler languages such as query languages. Thus almost any business-small, medium or large scale began using computers for day-today activities. Now what is the use of all this data? Up to the early 1990's the answer to this was "NOT much". No one was really interested in utilizing data, which was accumulated during the process of daily activities. As a result a new discipline in computer science, Data Mining gradually evolved.

Index Terms *– Data Mining, Applications, Trends*

DATA MINING

Data Mining is the exploration and analysis of large sets, in order to discover meaningful patterns and rules. The key idea is to find effective ways to combine computers power to process data with the human eye's ability to detect patterns. The techniques of data mining are designed for work best with large data sets.

Since Data Mining is a young discipline with wide and diverse applications, there is still a nontrivial gap between
general principals of Data Mining and domain specific, effective Data Mining tools for particular applications. In
this paper we will illustrate a few application domains of
Data Mining (such as finance, the retail industry and
telecommunication) and Trends in Data Mining which include further efforts towards the exploration of new application areas and new methods for handling complex
data types, algorithms scalability, constraint based mining
and visualization methods, the integration of data mining with data warehousing and database systems, the standardization of data mining languages, and data privacy protection and security.

INTRODUCTION

Data mining is the process of extraction of interesting (nontrivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. It is the set of activities used to find new, hidden or unexpected patterns in data or unusual patterns in data. Using information contained within data warehouse, data mining can often provide answers to questions about an organization that a decision maker has previously not thought to ask.

- Which products should be promoted to a particular customer?
- What is the probability that a certain customer will respond to a planned promotion?
- Which securities will be most profitable to buy or sell during the next trading session?
- What is the likelihood that a certain customer will default or pay back a schedule?
- What is the appropriate medical diagnosis for this patient?

These types of questions can be answered surprisingly easily if the information hidden among the petabytes of data in your databases can be located and utilized. In the

following paragraphs we will discuss about the applications and trends in the fields of data mining.

APPLICATIONS

As data mining matures, new and increasingly innovative
applications for it emerge. Although a wide variety of data mining scenarios can be described. For the purpose of this paper the applications of data mining are divided in the following categories:

- Healthcare
- Finance
- Retail industry
- Telecommunication
- Text Mining & Web Mining
- Higher Education

Healthcare

The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and in cancer therapies to the identification and study of human genome by discovering large scale sequencing patterns and gene functions. Recent research in DNA analysis has led to the discovery of genetic causes for many diseases and disabilities as well as approaches for disease diagnosis, prevention and treatment.

Finance

Most banks and financial institutions offer a wide variety of banking services (such as checking, saving, and business and individual customer transactions), credit (such as business, mortgage, and automobile loans), and investment services (such as mutual funds). Some also offer insurance services and stock services. Financial data collected in the banking and financial industry is often relatively complete, reliable and high quality, which facilitates systematic data analysis and data mining. For example it can also help in fraud detection by detecting a group of people who stage accidents to collect on insurance money.

Retail Industry

Retail industry collects huge amount of data on sales, customer shopping history, goods transportation and consumption and service records and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability and popularity of the business conducted on web, or e-commerce. Retail industry provides a rich source for data mining. Retail data mining can help identify customer behavior, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios design more effective goods transportation and distribution policies and reduce the cost of business.

Telecommunication

The telecommunication industry has quickly evolved from offering local and long distance telephone services to provide many other comprehensive communication services including voice, fax, pager, cellular phone, images, e-mail, computer and web data transmission and other data traffic. The integration of telecommunication, computer network, Internet and numerous other means of communication and computing are underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand from data mining in order to help understand business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service

Text Mining and Web Mining

Text mining is the process of searching large volumes of documents from certain keywords or key phrases. By searching literally thousands of documents various relationships between the documents can be established. Using text mining however, we can easily derive certain patterns in the comments that may help identify a common set of customer perceptions not captured by the other survey questions. An extension of text mining is web mining. Web mining is an exciting new field that integrates data and text mining within a website. It enhances the web site with intelligent behavior, such as suggesting related links or recommending new products to the consumer. Web mining is especially exciting because it enables tasks that were previously difficult to implement. They can be configured to monitor and gather data from a wide variety of locations and can analyze the data across one or multiple sites. For example the search engines work on the principle of data mining.

Higher Education

An important challenge that higher education faces today is predicting paths of students and alumni. Which student will enroll in particular course programs? Who will need additional assistance in order to graduate? Meanwhile additional issues, enrollment management and time-to degree, continue to exert pressure on colleges to search for new and faster solutions. Institutions can better address these students and alumni through the analysis and presentation of data. Data mining has

quickly emerged as a highly desirable tool for using current reporting capabilities to uncover and understand hidden patterns in vast databases.

TRENDS

As different types of data are available for data mining tasks, so data mining approaches poses many challenging
research issues in data mining. The design of a standard data mining languages, the development of effective and
efficient data mining methods and systems, the construction of interactive and integrated data mining environments, and the applications of data mining to solve large applications large application problems are important tasks for data mining researches and data mining system and application developers. Here we will discuss some of the trends in data mining that reflect the pursuit of these challenges:

Application Exploration

Earlier data mining was mainly used for business purpose, to overcome the competitors. But as data mining is becoming more popular it is gaining wide acceptance in other fields also such as biomedicine, stock market, fraud
detection, telecommunication and many more. And many new explorations are being done for this purpose. In addition for data mining for business continues to expand as e-commerce and marketing becomes mainstream elements of the retail industry.

Scalable data mining methods

The current data mining methods capable of handling only a particular type of data and limited amount of data, but as data is expanding at a massive rate, there is a need to develop new data mining methods which are scalable and can handle different types of data and large volume of data The data mining methods should be more interactive and user friendly. One important direction towards improving the repair efficiency of the timing process while increasing user interaction is constraint-based mining. This provide user with more control by allowing the specification and use of constraints to guide data mining systems in their search for interesting patterns.

Combination of data mining with database systems, data
warehouse systems, and web database systems, Database systems, data warehouse systems, and WWW are loaded with huge amounts of data and have thus become the major information processing systems. It is important to make sure that data mining serves as essential data analysis component that can be easily included in to such an information-processing environment. The desired architecture for data mining system is the tight coupling
with database and data warehouse systems. Transaction
management query processing, online analytical processing and online analytical mining should be integrated into one unified framework.

Standardization of data mining language:

Today few data mining languages are commercially available in the market like Microsoft's SQL server 2005,
IBM Intelligent Miner, SAS Enterprise Miner, SGI Mineset, Clementine , DBMiner and many more but a standard data mining language or other standardization efforts will provide the orderly development of data mining solutions, improved interpretability among multiple data mining systems and functions.

Visual data mining

It is rightly said a picture is worth a thousand words. So if
the result of the mined data can be shown in the visual form it will further enhance the worth of the mined data. Visual data mining is an effective way to discover knowledge from huge amounts of data. The systematic study and development of visual data mining techniques will promote the use for data mining analysis.

New methods for mining complex types of data

The complex types of data like geospatial, multimedia, time series, sequence and text data poses an important research area in field of data mining. There is still a huge gap between the needs for these applications and the available technology.

Web mining

The World Wide Web is huge collection of globally distributed collection of news, advertisements, consumer
records, financial, education, government, e-commerce and many other services. The WWW also contains huge and dynamic collection hyper linked information, providing a huge source for data mining. Based on the above facts, the web also poses great challenges for efficient resource and knowledge discovery.

Need of Data Mining

The massive growth of data from terabytes to petabytes is
due to the wide availability of data in automated form from various sources as WWW, Business,

science, Society and many more. But we are drowning in data but deficient of knowledge Data is useless, if it cannot deliver knowledge. That is why data mining is gaining wide acceptance in today's world. A lot has been done in this field and lot more need to be done.

Conclusion

Since data mining is a young discipline with wide and diverse applications, there is still a nontrivial gap between
general principles of data mining and domain specific, effective data mining tools for particular applications. A few application domains of Data Mining (such as finance, the retail industry and telecommunication) and Trends in Data Mining which include further efforts towards the exploration of new application areas and new methods for
handling complex data types, algorithms scalability, constraint based mining and visualization methods, the integration of data mining with data warehousing and database systems, the standardization of data mining languages, and data privacy protection and security.

References

[1] Data Mining Concepts and Techniques – Jiawei Han & Micheline Kamber
[2] Modern Data Warehousing, Mining and Visualization Core Concepts by George M.arakas.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Cloud Computing with Related Enabling Technologies

P.Rizwan Ahmed
Head & Asst. Professor
Department of Computer Applications and PG Department of Information Technology
Mazharul Uloom College, Ambur – 635802, Tamil Nadu
prof.rizwanphd@gmail.com

ABSTRACT

Cloud Computing is a concept that has been defined differently by many and there seem not to be a consensus. Despite these views, cloud computing is not a complete new idea as it has intricate connections to technologies or domain such as the Grid Computing paradigm, and the general distributed computing. This overview gives the basic concept of cloud computing, and highlights the relationship between Cloud computing and other cloud enabling technologies by providing their similarities and differences. This insight into the essential characteristics of cloud and its enabling technologies provides a good foundation for understanding and a hint on how to leverage desirable strengths of these technologies in the cloud by way of extension and or inheritance.

1.0 INTRODUCTION

Several technologies are related to cloud computing, and the cloud has emerged as a convergence of several computing trends. It seeks to address certain key aspects that may have been lacking in each of these trends, individually. In this paper, the features of each of these related technologies and how they compare and or relate with cloud computing is discussed.

Cloud computing is a paradigm shift to computing [1] that sees and delivers computing as a service rather than as a resource. Cloud computing encapsulates several layers of computing provisioning that include the hardware resources located at the data centers of cloud providers, the operating system and virtualization software on top of that hardware, and the applications that are delivered as services over the internet. These services are provided as utility to customers who are billed based on usage, similar to the billing scheme of traditional public services such as electricity, telephone and water.

Cloud Computing is a realization of computing pioneer John McCarthy‟s prediction back in 1961, that “computation may someday be organized as a public utility”; and went on to speculate how this might occur [2]. Cloud computing is not unconnected with Grid Computing and other related technologies.

In the mid 1990s, the term Grid was coined to describe technologies that would allow consumers to obtain computing power on demand. Ian Foster and others [3],[4] posited that by standardizing the protocols used to request computing power, we could spur the creation of a Computing Grid, analogous in form and utility to the

electric power grid. Researchers subsequently developed these ideas in many exciting ways, producing for example large-scale federated systems (TeraGrid, Open Science Grid, caBIG, EGEE, Earth System Grid) that provide not just computing power, but also data and software, on demand. Standards organizations (e.g., OGF, OASIS) defined relevant standards. This term was co-opted by industry as a marketing term for clusters. But no viable
commercial Grid Computing providers emerged, at least not until recently [2].

This article aims to explain the key concepts, and underlying technologies as well as discussing the relationships in terms of similarities and differences of cloud computing to other domain which has contributed to the evolution of cloud computing. This will help deliver a much broader understanding and appreciation of cloud computing.

The rest of this paper is organized as follows: Section 2 discusses the theoretical baselines,

section 3 provides an overview of cloud computing, section 4 discusses cloud computing related technologies which are regarded as enabling technologies of the cloud, section 5 summarizes and section 6 concludes.

2.0 THEORETICAL BASELINES

Cloud computing today is the beginning of "network based computing" over Internet in force. It is the technology of the decade and the beginning to the end of the dominance of desktop computing such as that with the Windows. It is also the beginning of a new Internet based service economy: the Internet centric, Web based, on demand, Cloud applications and computing economy [5].

2.1 Distributed Computing

Distributed computing refers to the very idea of using distributed systems that are generally multiple computers connected to each other via computer networks to collaboratively process a common goal. Those computers communicating can be homogeneous or heterogeneous, distributed globally or locally. According to the characteristics of localization or equality, distributed systems have different subsets, such as supercomputers, grids, clusters, web 2.0 and clouds [2].

Before going further into the subsets of distributed computing, an illustration is provided to visualize the interconnection between the concepts that will be explained in the following:

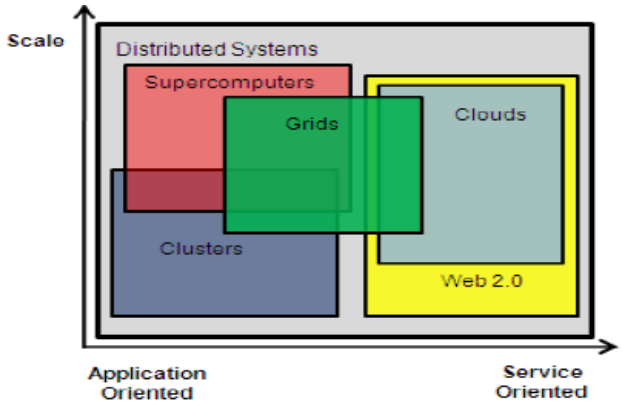


Figure 2.0: Distributed Computing and its subsets [2].

Every participanting machine in a distributed system is able to download the peace of software and then interconnect to a centralized server. The servers provide the input from their sensors with a huge amount of data, making the calculation of this data normally very complicated. While a single computer would not be able to do that in an appropriate amount of time, millions of computers that are interconnected anywhere in the world would be able to achieve that.

2.1.1 Clusters

Characteristics of clusters are that the computers being linked to each other are normally distributed locally, and have the same kind of hardware and operating system. Therefore cluster work stations are connected together and can possibly be used as a super computer [21].

2.1.2 Supercomputers

Supercomputers can be easily compared to clusters, because it follows the same concept, except the fact that it is merged into one box already and is not locally interconnected with other machines [21]. IBM is constructing those machines consisting with a lot of processors that are merged into one machine with high performance capabilities [22]. The only disadvantage is that they are usually expensive and have the necessity of a huge amount of energy.

2.1.3 Grids

When defining grid computing it is necessary to differ it from clusters. While clusters are distributed locally and obliged to use the same hardware and Operating System (OS), grids involve heterogeneous computers that are connected to each other and distributed globally. The OS and hardware that run on those machines can also be different from each other [21].

The computers that are interconnected over the internet can come from anywhere while there is usually no obligation to pay. For this reason already it is obvious that grids being connected are not nearly as expensive as the supercomputers that are offered from IBM and other technology companies.

2.1.4 Clouds

Together with virtualization, clouds can be defined as computers that are networked anywhere in the world with the availability of paying for the used clouds in a pay-per-use way, meaning that just the resources that are being used will be paid for [18]. In the following, clouds will be introduced.

3. CLOUD COMPUTING OVERVIEW

There is no absolute consensus on the meaning of the term "Cloud computing" as noted by [23],

"*A lot of people are jumping on the [cloud] bandwagon, but I have not heard two people say the same thing about it. There are multiple definitions out there of „the cloud.* "

Hence, many definitions have been given by different researchers to the term. We shall adopt the

America National Institute of Standards and Technology (NIST)"s definition of the cloud. According to NIST [6], Cloud computing is *"A pay-per-use model for enabling available convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."*

3.1 Cloud Computing Features

The summary of the features of Cloud Computing described by [7], [9] is:

- Cloud Computing is a new computing paradigm.
- Infrastructure resources (hardware, storage and system software) and applications are provided in everthing-as-a-service (XaaS) manner. When these services are offered by an independent provider or to external customers, Cloud Computing is based on pay-per-use business models.
- Main features of Clouds are virtualization and dynamic scalability on demand.
- Utility computing and SaaS are provided in an integrated manner, even though utility computing might be consumed separately.
- Cloud services are consumed either via Web browser or via a defined API.

3.2 Cloud Computing Characteristics

There are five essential characteristics of Cloud Computing which explains their relation and difference from the traditional computing. These are;

- *On-demand-self-service:* Consumer can provision or un-provision the services when needed, without the human interaction with the service provider.
- *Broad Network Access:* It has capabilities over the network and accessed through standard mechanism.
- *Resource Pooling (Multi-tenancy):* The computing resources of the provider are pooled to serve multiple consumers which are using a multi-tenant model, with various physical and virtual resources dynamically assigned, depending on consumer demand.
- *Rapid Elasticity:* Services can be rapidly and elastically provisioned.
- *Measured Service:* Cloud Computing systems automatically control and optimize resource usage by providing a metering capability to the type of services (e.g. storage, processing, bandwidth, or active user accounts) [7], [8].

3.3 Cloud Service Models

Three Cloud Services Models are defined, these fundamental classifications are often referred to as "SPI model" that is; software, platform or infrastructure as a service.

- *Cloud Software as Service:* This is a capability in which the consumer can use the provider' s applications running on the cloud.
- *Cloud Platform as Service:* In this type of service, the consumer can deploy the consumer created or acquired applications created by using programming languages or tools provided by provider, on the cloud infrastructure.
- *Cloud Infrastructure as Service:* This is a capability provided to the consumer by which, it can provision processing, storage, networks and other fundamental computing resources where the consumers can deploy and run the software (that is, operating systems, applications) [7], [8].

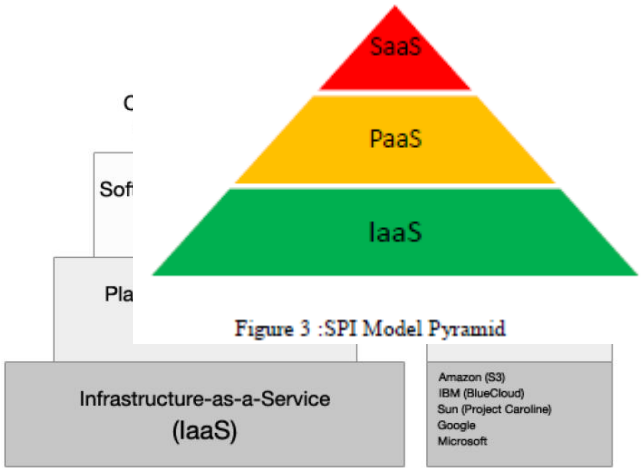


Figure 3 :SPI Model Pyramid

Figure 3.1: Showing the Cloud Computing Stack and Existing Cloud Services [20].

3.4 Cloud Deployment Models

Cloud services are provisioned in different ways, this are referred to as deployment models [7]. They include;

- *Public Cloud:* The cloud infrastructure is available to the general public.
- *Private Cloud:* The type of the cloud, that is available solely for a single organization.
- *Community Cloud:* In this type of cloud deployment model, the infrastructure of the cloud is shared by several organizations and supports a specific community with shared concerns.
- *Hybrid Cloud:* This is a cloud infrastructure that is a composition of two or more clouds that is, private, community or public [7], [8].

4. CLOUD COMPUTING RELATED TECHNOLOGIES

Several technologies are related to cloud computing, and the cloud has emerged as a convergence of several computing trends. It seeks

to address certain key aspects that may have been lacking in each of these trends, individually.

Clouds extend the capabilities of other domains with the specific goal to achieve scalability/ elasticity, availability with optimal resource utilization, which is as such only partially addressed in other domains. Specifically, clouds belong to the wider areas of Internet of Services (including Web Services, Web3.0, Service Oriented Architecture (SOA) etc.) and Utility Computing (including Grid, Virtual Organizations etc.) and implicitly inherit multiple aspects from these domains, such as virtualization and outsourcing. Depending on usage, they may extend these characteristics, in particular by adding a new business model. What is important to stress again in this context is that not all cloud characteristics exclusively belong to the cloud domain; extrinsic features that generally belong to other domains enable service and utility computing are naturally taken over in clouds. Specifically, we can distinguish (Figure 4.0) between

- Characteristics exclusive to the Cloud (intrinsic features)
- Characteristics belonging to other domains but having to be adapted in order to meet the cloud relevant specifics (extrinsic extended)
- Characteristics that belong to other domains and just act as enablers to cloud systems, that is, do not have to be extended (extrinsic inherited)

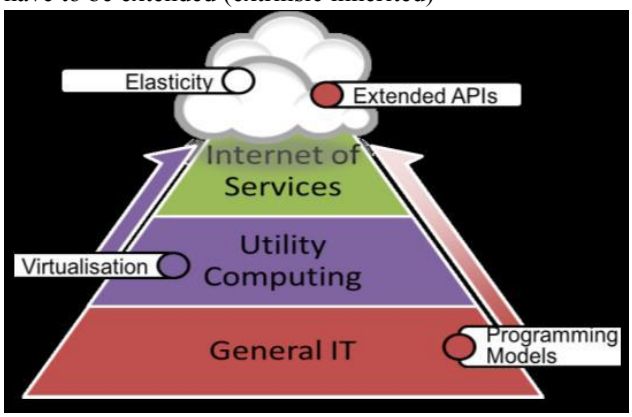


Figure 4.0: Inheritance and Extension of Characteristics across the Related Domains [10]

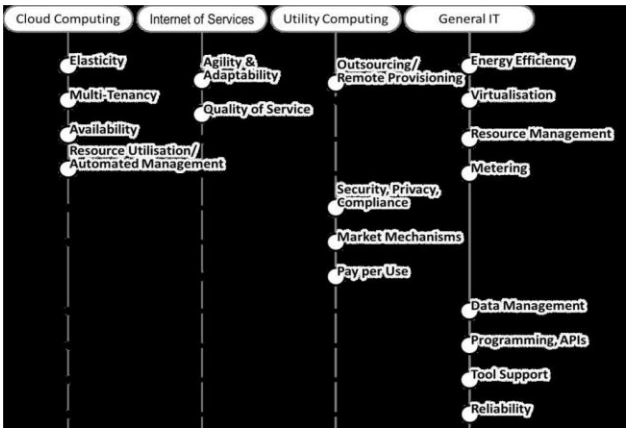


Figure 4.1: Overview of the Cloud Characteristics and their Relationship to other Domains. Inheritance is from Right to Left. The Further Left is Characteristics Advances, the more it needs to be adapted for the respective domain(s), [10].

We refer here in particular to three related domains, though the list can be easily extended and refined:

- Internet of Service: covering areas such as Web Services, Web 3.0, and SOA etc. In other words, individual service provisioning without specific means for dealing with availability, that is, just network load balancing.

- Utility Computing: including Grids, Virtual Organizations, and also High Performance Computing (HPC) to some degree, as they were originally conceived and realized.
- General IT or more correctly "non-web" IT, including all computer science aspects concerning isolated machines, i.e. without making explicit use of the web. This includes theory of computation, hardware architectures, operating systems etc.

More specifically, the characteristics identified above can be classified with respect to the domains they were originally conceived in and to the ones which take up/extend them (see Figure 4.1). The Columns denote the domain to which the concepts apply, so that any entry within an according column implies that the respective characteristic is specifically adjusted and or extended to meet the domain's requirements. In other words, single column entries imply that all domains to the left inherit the characteristics without significant adaptations, whilst domains to the right do not support this characteristic. If an entry spans multiple columns, it means that the base concept is [10] conceived in the right-most domain and that all domains to the left adapt the concept to their respective needs.

It will be noted that this classification is subject to many discussions due to the overlap in terminologies across domains. It is worth mentioning in this context that the classification is oriented towards the left, that is, "which domain provides capabilities that can be exploited on higher-level domains", but there is a noticeable right-orientation, too, where for example utility computing is improved by availability methodologies of the Cloud, without necessarily turning the respective domain into a Cloud system.

4.1 Cloud Computing Vs. Grid Computing

There has always been a debate about the evolution of Cloud Computing and the most important point in that is Grid Computing. Some

people call Cloud Computing and Grid Computing the same phenomena while others call Cloud Computing an extension of Grid computing. To find out the truth we need to know about Grid computing [7], [9].

Grid Computing is a complex phenomenon which has evolved through earlier developments in parallel, distributed and high performance computing (HPC) [12], [11]. One of the most cited definitions of Grid computing was from [1]

"A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities." [9]
After that, the development of support for generic IT resource sharing started to be measured as the real Grid problem; according to Foster,
"The real and specific problem that underlies the Grid concept is coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations. The sharing that we are concerned with is not primarily file exchange but rather direct access to computers, software, data, and other resources, as is required by a range of collaborative problem-solving and resource brokering strategies emerging in industry, science, and engineering" [3].

Virtual organizations in this definition can be defined as the dynamic group of individuals, groups, or organization who define the conditions and rules for sharing resources [13]. Some of the organizations have also defined the Grid computing with respect to the features. According to IBM,

"Grid computing allows you to unite pools of servers, storage systems, and networks into a single large system so you can deliver the power of multiple-systems resources to a single user point for a specific purpose. To a user, data file, or an application, the system appears to be a single enormous virtual computing system." [14].

4.2 Virtualization in Cloud Computing
The description of Cloud computing earlier and that of Grid computing here shows that Cloud computing and Grid computing have many similarities but also differ in many respects. The most differentiating concept is the use of virtualization in cloud giving rise to multi-tenancy. Cloud computing technologies could never exist without the use of the underlying technology known as Virtualization [20].

As presented in the Table 1, what makes Cloud computing different from Grid computing is "virtualization". Cloud computing leverages virtualization to maximize the computing power; Virtualization, by separating the logical from the physical, resolves some of the challenges faced by Grid computing [7], [15]. While Grid computing achieves high utilization by the allocation of multiple servers onto a single task or job, the virtualization of servers in Cloud Computing achieves high utilization by allowing one server to compute several tasks concurrently [7], [16].

With virtualization, applications and infrastructure are independent, allowing servers to be easily shared by many applications where applications are running virtually anywhere in the world. This is possible as long as the application is virtualized [18]. Virtualizing the application for the cloud means to package the bits of the application with everything it needs to run, including pieces such as a database, a middleware and an operating system; this self-contained unit of virtualized application can then run anywhere in the world [18]. Virtualization also allows building so-called sandboxes. Sandboxes assure a higher degree of security and reliability by providing a mechanism to run programs safely. It is commonly used to "execute untested code or programs from unverified third-parties, suppliers and untrusted users" [19].

Along with the differences in technology among Grid computing and Cloud computing, usage patterns are also different between them. Grid is usually used for job execution while clouds are more frequently used to support long-running services [17]. As mentioned above, there is a debate in the technology world that Cloud computing has evolved from Grid computing and that Grid computing is the foundation for Cloud computing, [2] for example describe the relationship between Grid and Cloud computing as follows:

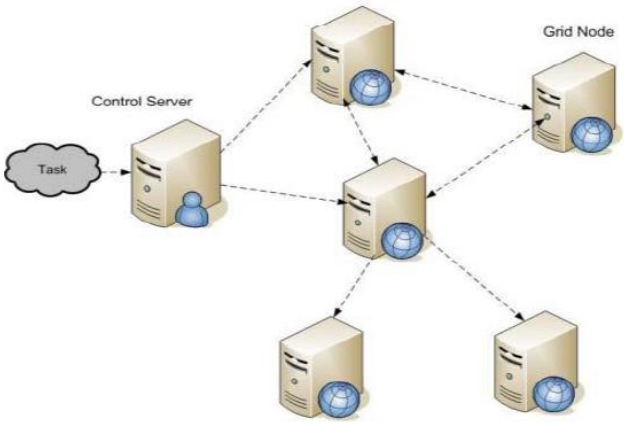
"We argue that Cloud Computing not only overlaps with Grid Computing, it is indeed evolved out of Grid Computing and relies on Grid Computing as its backbone and infrastructure support. The evolution has been a result of a shift in focus from an infrastructure that delivers storage and compute resources (such is the case in Grids) to one that is economy based aiming to deliver more abstract resources and services (such is the case in Clouds)." [2].

Another observed remarkable difference between cloud and grid computing is cloud's broad network access feature. This is depicted in figure 4.2 and figure 4.3 respectively. Other differences include the business model where the business model for

Grids (at least that found in academia or government labs)

is project-oriented in which the users or community represented by that proposal have certain number of service units (that is, CPU hours) they can spend while Cloud utilizes a pay-as-you-go business model.

It is worth noting that in the architecture of these two technologies, Cloud defines a four-layed architecture while Grid defines a five-layered architecture but both technologies have the same base and topmost layer (being the fabric as the base and application as the topmost layer). This makes the security model in Grid computing more



stringent and more secured, the reason is that security is provided by the protocol at each of the layers in the architecture thus, security has been engineered in the fundamental Grid infrastructure compared to cloud's security model which seems to be relatively simpler and less secure than the security model adopted by Grids. Cloud infrastructure typically rely on Web forms (over SSL) to create and manage account information for end-users, and allows users to reset their passwords and receive new passwords via Emails in an unsafe and unencrypted communication.

Figure 4.2: Grid Computing [5].

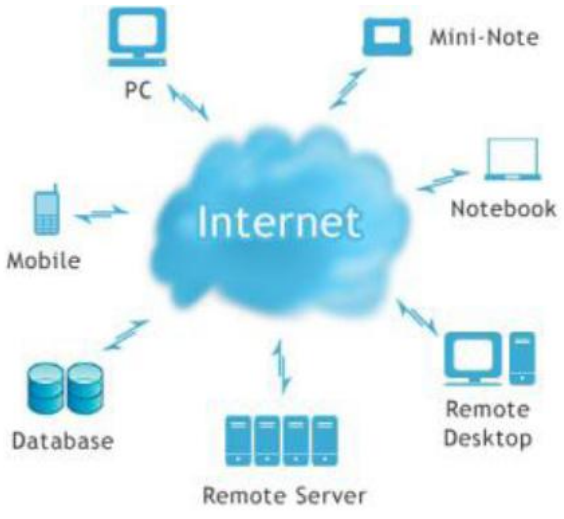


Figure 4.3: Cloud Computing [5].

5. SUMMARY

The vision of Cloud computing and Grid computing is the same [2] and that is to (i) reduce the cost of computing, (ii) increase reliability, and (iii) increase flexibility by transforming computers from something that we buy and operate ourselves to something that is operated by a third party even though the scale are different; clouds are operating at a different scale, and operating at these new, more massive scales can demand fundamentally different approaches to tackling problems.

The problems are mostly the same in Clouds and Grids. There is a common need to be able to manage large facilities; to define methods by which consumers discover, request, and use resources provided by the central facilities; and to implement the often highly parallel computations that execute on those resources. Details differ, but the two communities are struggling with many of the same issues. Thus we can summarize that Grid computing is the starting point and basis for Cloud computing. Cloud computing essentially represents the increasing trend towards the external deployment of IT resources, such as computational power, storage or business applications, and obtaining them as services [9].

6. CONCLUSION

In information technology, it is observed that technology scales by an order of magnitude, and in the process reinvents itself. Such is the case with Cloud computing. Cloud computing is a paradigm shift in computing that completely realized the predictions of John McCarthy. It inherited and also extended the features of preceding technologies such as Grid computing. Thus, Cloud computing possess to a high degree, features of the Grid but it is still not with some differences. One of these outstanding differences between Cloud and Grid is virtualization which has given rise to multi-tenancy in Cloud computing environment. It is believed at this advocacy level of Cloud computing that adopting key strengths of Cloud enabling technologies such as the Grid especially in areas of its security model will help advance the adoption of cloud computing.

REFERENCES

[1] Voas, J., & Zhang, J. (2009). Cloud Computing: New Wine or Just a New Bottle? *IT Professional*, 15-17.

[2] Foster I. et al. (2008). Cloud Computing and Grid Computing 360-Degree Compared. IEEE Grid Computing Environment Workshop (GCE „08), pp 1-10.

[3] I. Foster, C. Kesselman, S. Tuecke. (2001). The Anatomy of the Grid: Enabling scalable virtual organization. The International Journal of High Performance Computing Applications, vol. 15(3), pp. 200-222.

[4] I. Foster, C. Kesselman, J. Nick, S. Tuecke (2002). The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Globus Project.

[5] Srinivasa et al. ().Cloud Computing: An Overview. Journal of Theoretical and Applied Information Technology, vol. pp. 71-79. www.jatit.org.

[6] NIST (2010) NIST Definition of Cloud Computing v15. US Department of Commerce, Washington.

[7] Rehan S. (2011) "Cloud Computing‟s Effect on Enterprises", Dept of Informatics, Lund University Unpublished Master‟s Thesis 1-89.

[8] Cloud Security Alliance. (2009). *Security Guidance for Critical Areas of Focus in Cloud Computing*.

[9] Stanoevska-Slabeva, K., Wozniak, T. (2009). *Grid Basics. In: Stanoevska-Slabeva, K., Wozniak, T., and Ristol, S., Grid and Cloud Computing A Business Perspective on Technology and Applications*. Springer Berlin Heidelberg, 2009.

[10] Lutz S & Keithe J. (2012): Advances in Clouds: Research in Future Cloud Computing, Expert Group Report. Information Society and Media, European Union, Belgium.

[11] Harms, U., Rehm, H-J., Rueter, T., Wittmann, H. (2006) *Grid Computing für virtualisierte Infrastrukturen.* In: Barth T, Schüll A (eds) Grid Computing: Konzepte, Technologien, Anwendungen, pp. 1-15. ieweg+Teubner, Wiesbaden.

[12] Weishäupl T., Donno F., Schikuta E., Stockinger H., Wanek H. (2005). *Business In the Grid: The BIG Project.* In: Proceedings of the 2nd International Workshop on Grid Economics and Business Models (GECON 2005). http://hst.home.cern.ch/hst/publications/gecon-2005-BIGproject.pdf. Accessed 5th May, 2010.

[13] Joseph J, Ernest M, Fellenstein C (2004) *Evolution of Grid Computing Architecture and Grid Adoption Models.* IBM Syst. J. 43(4):624-644.

[14] Kourpas E (2006) *Grid Computing: Past, Present and Future – An Innovation Perspective.* IBM white paper.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Green Computing "Futures"

[1] P.Rizwan Ahmed, [2]A.Aqueel Ahmed
[1] Head & Asst. Professor , Department of Computer Applications and PG Department of Information Technology
[2]Asst. Professor, Department of Information System Management
Mazharul Uloom College, Ambur – 635802.

ABSTRACT

The concept of green computing has begun to spread in the past few years, gaining increasing popularity. Besides the widespread sensitivity to ecological issues, such interest also stems from economic needs, since both energy costs and electrical requirements of IT industry around the world show a continuously growing trend. Green computing is the environmentally responsible use of computers and related resources. Such practices include the implementation of energy-efficient central processing units (CPUs), Servers and Peripherals as well as reduced resource consumption and proper disposal of electronic waste (e-waste). Green computing is the study and practice of efficient and eco-friendly computing. The principle behind energy efficient coding is to save power by getting software to make less use of the hardware, rather than continuing to run the same code on hardware that uses less power. This paper, first discuss the connotation of green computing and sketch researcher's view on the next generation of IT systems for green computing. Subsequently, this paper helps to identify key issues relevant to green computing and evaluate different approaches to these problems. Finally, paper point out future directions of research and conclude the paper.

Index Terms – Energy Efficient Coding, Green Computing ,Green IT ,Smart Computing.

I. Introduction

Green computing, the study and practice of efficient and eco-friendly computing resources, is now under the attention of not only environmental organizations, but also businesses from other industries. In recent years, companies in the computer industry have come to realize that going green is in their best interest, both in terms of public relations and reduced costs. This article will take a look at several green initiatives currently under way in the computer industry, as well as issues that have been raised regarding these initiatives. We will also have a talk with VIA to learn more about the future of green computing. When we heard the term green computing the first thought that came into our mind was ―going green with computers‖ but the questions that strike our thoughts the very same moment were HOW and WHY to ―go green‖, and in the quest for finding the answers to our questions we landed up with the conclusion "GREEN COMPUTING –GREAT COMPUTING"[1]. The Green Computing Initiative, stewards of the industry standards EFGCD – Eco – Friendly Green Computing Definition defines Eco- Friendly Green Computing as the study and practice of the design, development, implementation, utilization and disposal of IT infrastructure efficiently and effectively with low or zero impact on the environment whilst reducing operating costs. Currently the ICT industry is responsible for 3% of the world's energy consumption. With the rate of consumption increasingly by 20% a year, 2030 will be the year when the world's energy consumption will double because of the ICT industry. Organizations use the Green Computing Lifecycle when designing and implementing green computing technologies. The stages in the Life Cycle include Strategy, Design, Implementation, Operations and Continual Improvements. The 5 core green computing technologies advocated by GCI are Green Data Center, Virtualization, Cloud Computing, Power Optimization and Grid Computing. Company like Via Technology offer green PC's that are affordable, non- toxic and ultra low wattage. It take responsibility for their outdated products by offering a PC recycling service. [2]Cutting back on these two energy uses - the computers themselves and the energy used to cool them - makes a direct impact on company costs. Cutting back on energy use by making things more efficient can bring secondary savings too, which may not be immediately obvious. If you can increase the energy efficiency of front and back-office computing, you may not need to increase hardware Research resources as quickly as you

thought, which can save capital expenditure on the kit, cooling equipment and even the buildings necessary to house them.[3]It provides managers, academicians, scientists, and researchers in various government, public, and private sectors coverage of topical issues like green strategy, green transformation, green technology, green revolution, ecology system, sustainability supply chain, green and sustainable innovation, global warming, energy efficient system, recycling and reuse systems, product usability, reverse supply chain, closed loop supply chain, environmental issues, carbon footprints, renewable energy, applied ergonomics, and climate change. This paper offers research contributions, constructive debates, and investigations on new legislations on green IT, green processes, healthcare informatics, and applications in terms of environmental and climate issues for both manufacturing and service industry

II. Literature Review

When it comes to PC disposal you need to know everything there is to know in order to be involved in green computing. Basically, the whole green aspect came about quite a few years back when the news that the environment was not a renewable resource really hit home and people started realizing that they had to do their part to protect the environment. [4] Many governments worldwide have initiated energy-management programs, such as Energy Star, an international standard for energy-efficient electronic equipment that was created by the United States Environmental Protection Agency in 1992 and has now been adopted by several other countries. Energy Star reduces the amount of energy consumed by a product by automatically switching it into ―sleep mode when not in use or reducing the amount of power used by a product when in ―standby mode. Surprisingly, standby ―leaking, the electricity consumed by appliances when they are switched off, can represent as much as 12 percent of a typical household‘s electricity consumption. Basically, the efficient use of computers and computing is what green computing is all about. The triple bottom line is what is important when it comes to anything green and the same goes for green computing. This considers social responsibility, economic viability and the impact on the environment. Many business simply focus on a bottom line, rather than a green triple bottom line, of economic viability when it comes to computers. The idea is to make the whole process surrounding computers more friendly to the environment, economy, and society. This means manufacturers create computers in a way that reflects the triple bottom line positively. Once computers are sold businesses or people use them in a green way by reducing power usage and disposing of them properly or recycling them. The idea is to make computers from beginning to end a green product. The solution to green computing is to create an efficient system that implements these factors in an environmentally friendly way. A good example would be IT managers purchasing hardware that has been EPEAT approved meaning that maintenance is reduced, the hardware's life is extended, and makes recycling the computer easy once it is no longer necessary. Mobile phones are better than computers – green computing. What do you use your computer for? Surfing Internet, chat, gaming, social networking, downloading, desktop computing including documents, spreadsheets or presentation making or just watching your photos and videos ? Today_s mobile phones are capable of doing it all, rather sometimes more than the traditional phones. They have faster processors, more ram, faster wireless Internet connectivity and larger memories. Mobile Phones consume very low power. VIA Technologies, a Taiwanese company that manufactures motherboard chipsets, CPUs, and other computer hardware, introduced its initiative for "green computing If everyone takes into account green computing then our world of computers will have as little a negative impact on our physical world as possible and that is what green computing is all about.

III. Technologies Green Computing

VIA Technologies, a Taiwanese company that manufactures motherboard chipsets, CPUs, and other computer hardware, introduced its initiative for "green computing" in 2001. With this green vision, the company has been focusing on power efficiency throughout the design and manufacturing process of its products. Its environmentally friendly products are manufactured using a range of clean-computing strategies, and the company is striving to educate markets on the benefits of green computing for the sake of the environment, as well as productivity and overall user experience.

A. Carbon-free computing

One of the VIA Technologies‘ ideas is to reduce the "carbon footprint" of users — the amount of greenhouse gases produced, measured in units of carbon dioxide (CO2). Greenhouse gases naturally blanket the Earth and are responsible for its more or less stable temperature. An increase in the concentration of the main greenhouse gases — carbon dioxide, methane, nitrous oxide, and fluorocarbons — is believed to be responsible for Earth's increasing temperature, which could lead to severe floods and droughts, rising sea levels, and other environmental effects, affecting both life and the world's economy

B. Solar Computing

Amid the international race toward alternative-energy sources, VIA is setting its eyes on the sun, and the company's Solar Computing initiative is a significant part of its green-computing projects. For that purpose, VIA partnered with Motech Industries, one of the largest producers of solar cells worldwide. Solar cells fit VIA's power-efficient silicon, platform, and system technologies and enable the company to develop fully solar-powered devices that are nonpolluting, silent, and highly reliable. Solar cells require very little maintenance throughout their lifetime, and once initial installation costs are covered, they provide energy at virtually no cost. Worldwide production of solar cells has increased rapidly over the last few years; and as more governments begin to recognize the benefits of solar power, and the development of photovoltaic technologies goes on, costs are expected to continue to decline. As part of VIA's ―pc-1‖ initiative, the company established the first-ever solar-powered cyber community center in the South Pacific, powered entirely by solar technology.

C. Lead-Free and RoHS computing

In February 2003, the European Union adopted the Restriction of Hazardous Substances Directive (RoHS). The legislation restricts the use of six hazardous materials in the manufacture of various types of electronic and electrical equipment. The directive is closely linked with the Waste Electrical and Electronic Equipment Directive (WEEE), which sets collection, recycling, and recovery targets for electrical goods and is part of a legislative initiative that aims to reduce the huge amounts of toxic e-waste. Driven by these directives, VIA implemented a set of internal regulations in order to develop products that are compliant with these accepted policies, including the use of nonhazardous materials in its production of chipsets, processors, and companion chips. In 2001, they focused on lead-free manufacturing, introducing the Enhanced Ball Grid Array (EBGA) package for power efficient VIA processors and the Heat Sink Ball Grid Array (HSBGA) package for their chipsets. In traditional manufacturing processes, lead is used to attach the silicon core to the inside of the package and to facilitate integration onto the motherboard through tiny solder balls on the underside of the package. VIA's lead-free manufacturing technologies do not require a lead bead, and the solder balls now consist of a tin, silver, and copper composite.

D. Energy-efficient computing

A central goal of VIA‘s green-computing initiative is the development of energy-efficient platforms for low-power, small-form-factor (SFF) computing devices. In 2005, the company introduced the VIA C7-M and VIA C7 processors that have a maximum power consumption of 20W at 2.0GHz and an average power consumption of 1W. These energy-efficient processors produce over four times less carbon during their operation and can be efficiently embedded in solar-powered devices.

VIA isn‘t the only company to address environmental concerns: Intel, the world's largest semiconductor maker, revealed eco-friendly products at a recent conference in London. The company uses virtualization software, a technique that enables Intel to combine several physical systems into a virtual machine that runs on a single, powerful base system, thus significantly reducing power consumption. Earlier this year, Intel joined Google, Microsoft, and other companies in the launch of the Climate Savers Computing Initiative that commits businesses to meet the Environmental Protection Agency‘s Energy Star guidelines for energy-efficient devices.

E. vision through the pc-1 initiative

VIA isn‘t focusing only on the technological aspects of its eco-friendly devices, it‘s also taking a look at their applications. The VIA pc-1 initiative seeks to enable the next 1 billion people to get connected, by providing wider access to computing and communications technologies. The company is concentrating on empowering new, emerging markets, looking at models that reach beyond individual ownership of a PC, such as local pay-for-use facilities. Products built for such a use are characterized by ultra-efficient energy consumption and the ability to withstand heat and dust in harsh environments. In VIA‘s own words: "*Pc-1* brings together business ingenuity with corporate responsibility and altruism. Helping to build skills and literacy throughout the world and incorporating and preserving cultural content are goals now within our grasp. Information is the oxygen to nurturing social mobility, economic equality and development, and global democracy. Providing not just the tools and the know-how, but the support and the maintenance, is all part of what makes *pc-1* the next generation of information technology, the next generation of global development". Among the company‘s projects under the pc-1 program are the ”tuXlab” computer center in South Africa and an ICT Training Center in Vietnam.

IV. WHY GREEN COMPUTING?

In a world where business is transacted 24/7 across every possible channel available, companies need to collect, store, track and analyze enormous volumes of data—everything from click stream data and event logs to mobile call records and more. But this all comes with a cost to both businesses and the environment. Data warehouses

and the sprawling data centers that house them use up a huge amount of power, both to run legions of servers and to cool them. Just how much? A whopping 61 billion kilowatt-hours of electricity, at an estimated cost of $4.5B annually. The IT industry has begun to address energy consumption in the data center through a variety of approaches including the use of more efficient cooling systems, virtualization, blade servers and storage area networks (SANs). But a fundamental challenge remains. As data volumes explode, traditional, appliance-centric data warehousing approaches can only continue to throw more hardware at the problem. This can quickly negate any green gains seen through better cooling or more tightly packed servers. To minimize their hardware footprint, organizations also need to shrink their "data footprint" by addressing how much server space and resources their information analysis requires in the first place. A combination of new database technologies expressly designed for analysis of massive quantities of data and affordable, resource-efficient, open-source software can help organizations save money and become greener. Organizations can do so in the following three key areas: reduced data footprint, reduced deployment resources, and reduced ongoing management and maintenance. This technology is beneficial as it:-

- Reduce energy consumption of computing resources during peak operation
- Save energy during idle operation
- Use eco-friendly sources of energy
- Reduce harmful effects of computing resources
- Reduce computing wastes

Global warming and the problem of minimizing environmental impact from fossil-fuel emissions have raised to the top of global public policy agenda. As a result, businesses and consumers alike have started to embrace environmentally sustainable products that offer low-carbon solutions that can not only reduce their global greenhouse gas (GHG) emissions, but can do so by more efficient energy consumption and lower costs.

V. Conclusion

The field of "green technology " encompasses a broad range of subjects — from new energy-generation techniques to the study of advanced materials to be used in our daily life. As part of the VIA Green Computing Initiative, VIA Carbon Free Computing is a natural extension of VIA's leadership in developing the most power efficient computing products on the market. As individuals and organizations around the world look to reduce their impact on the environment, a growing concern is the reduction of one's Carbon Footprint which is a measure of the impact human activities have on the environment in terms of the amount of green house gases produced, measured in units of carbon dioxide (CO_2). It has taken upon itself the goal to provide society‘s needs in ways that do not damage or deplete natural resources. Mainly this means creating fully recyclable products, reducing pollution, proposing alternative technologies in various fields, and creating a center of economic activity around technologies that benefit the environment. Green IT programs are demonstrating fundamental economic as well as environmental sense, it is understandable why organizations are exploring green computing options with such intense interest across the IT industry. As more and more companies include some form of reporting on their goals and achievements in the area of CSR, there is a growing awareness among business leaders that greening their IT practices offers the ―double-win‖ of reducing costs while demonstrating a positive environmental commitment. Use mobile phones for your computing needs whenever and wherever possible

VI. Findings

1. If we think computers are non polluting and consume very little energy we need to think again. It is estimated that out of $250 billion per year spent on powering computers worldwide only about 15% of that power is spent computing- the rest is wasted idling. Thus, energy saved on computer hardware and computing will equate tonnes of carbon emissions saved per year.

2. The plan towards green IT should include new electronic products and services with optimum efficiency and all possible options towards energy savings.

3. power supplies are notoriously bad, generally as little as 7% efficient. And since everything in a computer runs off the power supply, nothing can be efficient without a good power supply. Recent inventions of power supply are helping fix this by running at 80% efficiency or better

4. Mobile phones are better than computers – green computing. They have faster processors, more ram, faster wireless Internet connectivity and larger memories. Mobile Phones consume very low power

5. Purchase LCD‘s monitors which consume less energy than CRT‘s screen and LCD‘s is also not harmful for the eyes.

References

[1]. Priya Rana, International Journal of Advanced Computer and Mathematical Sciences December 2010- Green Computing Saves Green

[2]. Simon Williams, Business, November 1st, 2009 - Green Computing

[3]. K. Ganesh (McKinsey & Company, India), International Journal of Green Computing (IJGC) October 20th,2012- Reach Your Enviromental Goals With Green Computing

[4]. Ms. Swati Aggarwal, Mrs. Monika Garg, Mr. Pramod Kumar, International Journal of Emerging Technology and Advanced Engineering, February 2012- Green Computing is Smart Computing – A Survey

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.

International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Mobile Commerce and Related Mobile Security Issues

P.Rizwan Ahmed
Head & Asst. Professor
Department of Computer Applications and PG Department of Information Technology
Mazharul Uloom College, Ambur – 635802, Tamil Nadu

ABSTRACT

This article will discuss how m-commerce conducts transactions of the mobile device through Internet and how these technologies are developed throughout the years. The article will also judge the security and privacy levels when dealing with mobile commerce, and what kind of issues are encountered when using mobile commerce systems. The article will also evaluate the solutions on how m-commerce issues are avoided and how they are tackled by the technology evolution.

Index Terms – M-Commerce, Security and Privacy Issues, Mobile Devices, Network.

1. Introduction

The term e-commerce signifies business processes on the Internet, such as the buying and selling of goods and services. There is a distinction between B2B (business-to-business) and B2C (business-to consumers) markets. In the first case, the business process is carried out between businesses and in the other case it is carried out between business and end consumers.[7]

The Internet and all kind of telecommunications have been gradually more common in many daily life aspects since the 1990s. In 2005, there were about 964 million Internet users and 2,168 million mobile phone users globally (Internet Telecommunication Union (2007) [1], [2], [3].

The expansion of the Internet and related technologies moved us to a range of new opportunities for business, providing businesses with new ways to conduct trade and exchange and communicate information through the progress and expansion of the e-commerce market (OECD, 2001)[1], [3], [7]. Therefore, a new type of communication service using Internet before our eyes by mobile device, and that lead to the emerging of new opportunities and to carry out different digital contents or services, called **mobile commerce** (m-commerce) (SERI, 2003). Belonging only to a specific individual, a mobile device let users to access service more compactly as well as to have more modified services than those of the wired Internet. However, many features such as cell phone ring tone download, music streaming or download, photo and video file transfer and download, mobile game download, e-mail service, search service, video streaming, and GPS service, were introduced not long before mobile phones were created according to the frequency of uses. The most frequently used services are mostly related to make download or location-based service easer. The article will also cover some of the benefits of m-commerce and how we can improve it [8], [6], [12].

2. Benefits of mobile commerce.

a) Mobile commerce involves all kind of electronic transactions by the use of mobile phone.
b) By the use of mobile commerce enterprises can improve and widen their market reach, cut down on cost
and give customers better service
c) Users can benefit from m-commerce by the convenience the m-commerce provide to them and organizing personal data
d) Examples of m-commerce are mobile parking meter payments and buying ringtones and games online.
Those kind of services are recognized as Micropayments: any transaction cost lower than $10
e) High value purchases such as land, houses and cars will be more convenient in the future [11].

3. Issues that concern M-commerce in terms of security and privacy.

The most important element when we are dealing with m-commerce is security issues and how we can
make it safe for customers to feel comfortable when using mobile phones, so in order to attract as many
customers we need to insure the quality of the security level provided [3], [15].
However, it's absolutely crucial to insure the safety of all kind of m-commerce transaction special those

that involve money transactions, therefore there are three main arias that this article will cover and elaborate
[2].

A. Security issues related to the network technologies.

This article will first look at, the challenges that face the security network technologies issues [8],[5].

GSM stand for Global System Mobile commerce, GSM's are the most used and the most common in the world especially in Europe, they are also considered to be the most important element of mobile commerce now a day's [3]. GSM's where first produced in the 1990s starting with a small market and very limited because it was not very handy back then, and also because it was very slow and above all that the device would not run without a computer which made it very inconvenient for customers to use it everywhere [10], [15], [16], [20].

As technology developed GSM and cell phones became more and more popular and more services were provided, such as SMS, wireless application protocols, HSCSD, and GPRS [3],[15].

GSM is responsible for the security of the mobile station when linked to a network, examples [10], [15], [20]:

i. IMSI confidentiality.
ii. IMSI data confidentiality on physical connections.
iii. Connectionless user data confidentiality.
iv. Signaling information element confidentiality.

The Wireless Local Area Network, which is known as (WLAN), operates in the unlicensed 2.4GHz, and
most mobile phones have this function and it is also becoming very popular [20]. However during the default
mode WLAN is not secured which makes the device easy to corrupt, so a certain level of security was
needed and that is why the IEEE invented WEP (Wired Equivalent Privacy) in order to solve the following
problems:

i. Authentication to protect the association to an AP.
ii. Integrity protection to MAC frames.
iii. Confidentiality to MAC frames [2], [13], [19].

B. Layer Security Transport

In this part the article will look at end-to-end security issues that concern mobile devices. Layer security
transport gives more security to the wire Logic, by this technology SMS and Packet data service was improved [2], [15], [20].

4. Solutions on how to solve the main issues of mobile commerce.

i. Security transaction over the web browser: when a customer is using mobile transaction though a web browser the customer is protected by ***inactivity lock*** *out,* this technology loges out the user automatically when the connection is lost [9], [14], [17].

ii. ***USSD***: this technology will make sure that all kind of transaction been made does not fall on the wrong hands and it is very powerful [9] [14], [17].

5. Discussion

Mobile commerce information security and privacy issues are a very important fact, which needed to be considered by mobile, m-commerce and other electronic commerce developers (Miyazaki and Fernandez, 2001; Earp and Anton, 2004) [2].

Consumers are very aware of those issues; therefore it will directly have an influence the services
provided by business such as money transactions and other services (Malhotra et al, 2004; Brown and
Muchira, 2004; Sah and Han, 2003) [2].

Technology has significantly evaluated from 'E-decade' to 'M-decade' (Wagner, 2005) [2].

There are about three billion subscribers using mobile phones worldwide. On the other hand, there are only one billion users to the Internet, this enormous wide spread of mobile phones technologies and the significant number of mobile devices which is increasing rapidly, will provide more opportunities for mobile commerce [19].

Mobile commerce was first found in 1997 in Finland, it was enhanced in a vending machine to serve Coca-Cola by using SMS [18].

6. Conclusion

In conclusion, the article has looked at some of the most sensitive arias of the new mobile commerce
such as, bank transaction systems because it engages with large amounts of money every day. Therefore the system used for transactions has to absolutely secure and to be free of corruption as mentioned earlier in the article; business will lose customers if the mobile security system is not

secure enough [6]. On the other hand the article also discussed the variety of ways of how our mobile commerce works and how it is been improved through the years. It also showed the kind of protocols that have been developed in order to make our mobile commerce safe and secure.

7. REFERENCE

[1] Hua Dai, University of North Carolina at Greensboro, Lakshmi Iyer, The University of North Carolina at Greensboro, Rahul Singh, University of North Carolina, Greensboro [An Investigation of Consumer's Security and
Privacy Perceptions in Mobile Commerce], (2007).

[2] Mark S. Ackerman and Donald T. Davis, Jr. Privacy and Security Issues in E-Commerce, (2008).
[3] Ali Grami and Bernadette H. Schell, Faculty of Business and Information Technology University of Ontario Institute of Technology ali.grami@uoit.ca and bernadette.schell@uoit.ca, [Future Trends in Mobile Commerce: Service Offerings, Technological Advances and Security Challenges] (2002).

[4] **Peter Tarasewich,** *College of Computer Science Northeastern Univers*ity tarase@ccs.neu.edu. **Robert C. Nickerson,** *College of Business San Francisco State University.* **Merrill Warkentin,** *College of Business & Industry Mississippi State University.* (ISSUES IN MOBILE E-COMMERCE), Communications of the Association for Information Systems (Volume 8, 2002).

[5] Norman Sadeh, (M-Commerce Technologies, Services, and Business Models), (2002).

[6] Ravi Tandon , Swarup Mandal and Debashis Saha,
(M-Commerce-Issues and Challenges), (2009).

[7] **Sanwar ALI,** Indiana University of Pennsylvania Department of Computer Science Indiana, **Waleed FARAG,** Zagazig University Department of Computer and Systems Engineering Zagazig, Egypt, **Mohammad A. ROB,** Management Information Systems University of Houston- Clear Lake Houston, (Security Measures in Mobile Commerce: Problems and Solutions), (2001).

[8] Seema Nambiar, Chang-Tien Lu, *Department of Computer Science Virginia Polytechnic Institute and State*
*University, (*Analysis of Payment Transaction Security in Mobile Commerce*), (2004).*

[9] Stefan Weiss, (Security and Privacy in Mobile Commerce), (2006).

[10] Sanjeev Banzal, (Mobile Banking & M–Commerce and Related Issues), (2010).

[11] A. S. Andreou, C. Chrysostomou, C. Leonidou, S. Mavromoustakos, A. Pitsillides, G. Samaras, C. Schizas,
(MOBILE COMMERCE APPLICATIONS AND SERVICES: A DESIGN AND DEVELOPMENT APPROACH), (2003).

[12] Shintaro Okazaki College of Economics and Business Administration Autonomous University of Madrid, Spain, (NEW PERSPECTIVES ON M-COMMERCE RESEARCH), (2005). [13] COMMITTEE ON CONSUMER POLICY (CCP), July (2009).

[14] Scarlet Schwiderski-Grosche, Heiko Knospe. (secure M-commerce), (2000).
[15] Suresh Chari1, Parviz Kermani, Sean Smith, and Leandros Tassiulas, (Security Issues in M–Commerce: A Usage– Based Taxonomy), (2006).

[16] Niels Christian Juul and Niels Jorgensen, (Security Issues in Mobile Commerce Using WAP), (2002).

[17] Linck, K., Pousttchi, Key and Wiedemann, Dietmar Georg University of Augsburg, (Security Issues in Mobile
Payment from the Customer Viewpoint), (2006).

[18] (M-Commerce and its Security Issues) By SAMEER YADAV (2001).

[19] Journal of Theoretical and Applied Electronic Commerce Research, (Special Issue on M-Commerce), AUGUST (2007).

[20] (M-commerce security white paper key security techniques) by Diarmuid Malloon, version 1.1 march (2010).



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A Study of Mobile IP

P.Rizwan Ahmed
Head & Asst. Professor
Department of Computer Applications and PG Department of Information Technology
Mazharul Uloom College, Ambur – 635802, Tamil Nadu

ABSTRACT

Connectivity to the Internet while in motion is becoming an extremely important part of computing research and development. Mobile IP, created by the Internet Engineering Task Force (IETF), is a standard protocol that builds on Internet Protocol by making mobility of a user transparent to applications and higher-level protocols such as Transfer Control Protocol. Mobile IP can be seen as the least common mobility denominator – providing seamless macro mobility solutions among the diversity of access. This paper will attempt to introduce Mobile IP from a technical point of view, while taking into consideration that the reader may not know anything about Mobile IP. However, the reader should know some networking basics before reading further. Building on these concepts, this paper will then discuss effective implementations of Mobile IP, the protocols used by Mobile IP and the importance of Mobile IP. This paper will also introduce Mobile IP from a consumer perspective, i.e. electronic devices (and their Operating Systems) which allow networking mobility.

Index Terms – Mobile IP, Mobile IP is IP-within-IP, Mobile IP Protocols

1. INTRODUCTION

The Internet is an excellent source of information which is readily accessible from almost any computer with a fixed connection to some kind of a network, however with increasing popularity of mobile devices such as PDA's, internet ready cell phones, PC Tablets, etc, there is a need to provide access to the Internet from a device that may be constantly in motion or wireless access to the Internet. Mobile IP, a standard proposed by the Internet Engineering Task Force (IETF) aims to make mobile computing a reality. The principal advantage of Mobile IP is that it frees the user from a fixed location. Mobile IP makes invisible the boundaries between attachment points, it is able to track and deliver information to mobile devices without needing to change the device's long-term Internet Protocol (IP) address (for that session) [1.CP.2002].

Before studying Mobile IP, it is important to define the concept of 'computing mobility'. Computing mobility can be defined as allowing the user some degree of freedom for his/her computing tasks. There are essentially two kinds of computing mobility – Personal mobility and Terminal mobility [2.IP.2003].

Personal mobility involves making it possible for a user to use the network's services from any terminal. When the user logs onto a terminal, he/she will get the same functionality as the user's home network, without having to go through time-consuming configuration procedures. When this user starts an IP session, an IP address will be assigned if it has not been done so already. As long as the IP session is kept alive, the IP address will serve as a destination for information delivery. An inherent problem with this approach is that once logged on, the user cannot switch to any other terminal. To use another terminal the user will have to repeat the login process on that terminal. An example of this may be an employee's laptop. At work, when the laptop is docked, the IP address will serve for communications, however, if the employee were to work from home over a high-speed connection or dial-up, the IP address of the laptop will be different, but the destination network will be the same. This approach requires technologies such as Network Access Identifier (NAI), where there is a relationship between the user's IP address and the NAI.

Terminal mobility means that the user and the terminal are mobile as one entity. The terminal may change its point of attachment[1] with the home network without:

- Informing the network to which it is connected to
- Having any impact on ongoing network services

Terminal mobility is tied to the Mobile IP protocol itself. Mobile IP does these two things in a seamless and lossless manner, in fact, ideally, the user should not ever be aware of the processes in

[1] A Point of Attachment can be defined as the access point of an entity into a network.

the background which allow the mobility. This mobility is handled in the network layer[2]; hence application session continuity is inherently provided for, as they will not even have to deal with mobility.

Mobile IP faces many obstacles, but ingenuity in modifying existing technology and communication protocols and adding new technologies have made Mobile IP a reality. Once the technology is perfected, users will enjoy the convenience of seamless untethered roaming and application transparency of nomadic computing[3].

Though the Internet can seem anarchic, IP routing depends on a well-ordered hierarchy. At the Internet core, routers are not concerned with individual users. They look at only the first few bits of an IP address (the prefix) and forward the packet to the correct network. Once at the network, routers within the network further look at the next few bits on the IP address, and send the packet to a subnet. At the edge, access routers look at the final parts of an address and send the packet to a specific machine [3.AD.2002].

To summarize the features of Mobile IP:

- No geographical limitations
- No physical connection required
- Modifications to other mobile devices/routers is not required
- Mobile IP leaves transport and higher-level protocols unaffected
- No modifications to the current IP address of the mobile device or the format of the IP address
- Supports security or implements some kind of authentication scheme to provide security

2. HOW MOBILE IP WORKS

IP addresses are typically associated with a fixed non-mobile location such as a router or a client computer. IP routes packets from a source to a destination by allowing routers to forward packets from incoming network interfaces to outbound interfaces according to routing tables. These routing tables typically maintain the next-hop information for each destination IP address, which is based on the number of networks to which that IP address is connected. The network number is derived from the IP address by masking off some of the low order bits. Thus, the IP address specifies the node's[4] point of attachment.

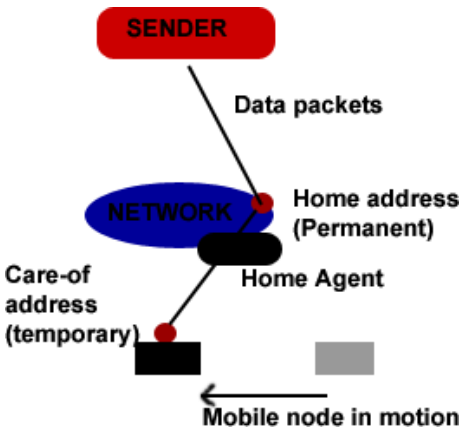
To maintain existing transport-layer connections the node must maintain a single IP address. In Transfer Control Protocol (TCP), the overwhelmingly popular protocol for Internet connections, the connections are indexed by a quadruplet that is analogous to someone's geographical home address. This set of numbers is what allows for delivery of a packet of data. If any of these numbers are changed then the connection will most likely be lost. Correct delivery of data packets to the node depends on the network number contained within the node's IP address.

If the node is mobile, packets sent to this node may never make it as, logically, mobility will force a new IP address to be associated with the node every time it changes its connection point. Constantly changing IP address of a node will make transparent mobility impossible.

Mobile IP was designed to solve this problem by allowing the node to use two IP addresses:

- Home address
- Care-of address

Figure 2.1
Simple diagram showing concept of having two IP addresses and a Home agent



Home address is static and is used to identify TCP connections. By definition, a home address is an address that makes the mobile node appear logically connected to the home network[5]. Mobile IP enables mobile nodes to surf the Internet, but remain addressable via their home network

[2] Discussion of Mobile IP in terms of network layers is done in section 4.

[3] Nomadic computing is analogous to Terminal Mobility.

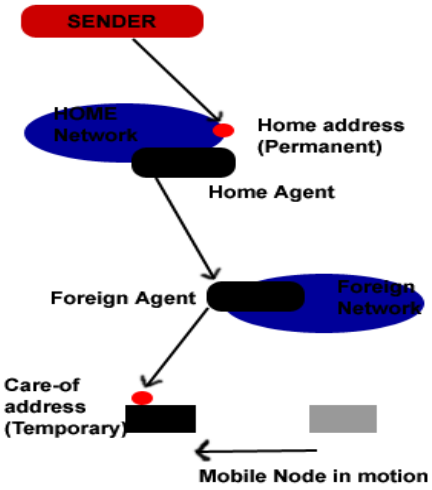
[4] '*node', 'client computer', and 'client'* are used interchangeably to mean a computer that is the receiving station.

[5] A Home Network of a mobile node is the network to which the mobile node is a part of, and thus to the rest of the Internet & other services.

[4.CP.1996]. The care-of address changes at each new point of attachment; it indicates the network number and thus identifies the mobile node's point of attachment with respect to the network topology.

The home address makes it appear that the mobile node is continually able to receive data on its home network where a network node known as the home agent assists in this operation. The home agent's principal job is to get data packets intended for the mobile node and deliver them transparently to the mobile node's current point of attachment. Whenever the mobile node is in a foreign network[6], delivery of data to the mobile node is taken care of by the foreign agent.

Figure 2.2
Mobile node on a Foreign Network



As mentioned, the care-of address changes at every new point of attachment. Whenever the point of attachment changes, the mobile node registers its new care-of address with its home agent. To get a packet to the mobile node from its home network, the home agent delivers the packet from the home network to the care-of address. To ensure delivery, the packet's header is modified so that the destination IP address is the care-of-address. This process can be thought of as a redirection where the packet is transformed from its original state to get delivered to its destination – the mobile node. This new header then shields or encapsulates the original packet, causing the mobile node's home address to have no effect on the encapsulated packet's routing until it arrives at the care-of address. Such encapsulation is also called tunneling[7]. This allows the packet to bypass the usual effects of IP routing.

The need for the existence of the home agent, foreign agent, and two different IP addresses is because of the lack of end-to-end network layer transparency, which is sometimes referred to as "Internet fog" [2.IP.2003]. Essentially, an IP address has to find its way through the Internet in order to get to its destination with the help of routing devices on the way. For Mobile IP to work properly, functionality has to be built-in so that it can find a way through the "fog".

So, the basic entities of a network that supports Mobile IP are:

- The mobile node
- The foreign agent
- The home agent
- The sender or correspondent node

Mobile IP can be understood as the cooperation of 3 separate mechanisms [5.CP.1999]:

2.1 Discovering the care-of address
2.2 Registering the care-of address
2.3 Tunneling to the care-of address

2.1 Discovering the care-of address

The Mobile IP discovery process has been built on top of an existing standard protocol, Router Advertisement. Internet Control Message Protocol (ICMP) router discovery messages are known as Router Advertisement. Router Advertisement enables hosts attached to multicast or broadcast networks to discover the IP address of their neighboring routers. Each router periodically multicasts a Router Advertisement from each of its multicast interfaces, announcing the IP address of that interface. Hosts discover the addresses of their neighboring routers by simply listening for advertisements [6.SED.1991].

Mobile IP does not modify the original fields of existing router advertisements but simply extends them to associate mobility functions. Thus, a router advertisement can carry information about default routers, just as before, and in addition carry further information about one or more care-of addresses. When the router advertisements are extended to also contain the needed care-of address, they are known as agent advertisements. Home agents and foreign agents typically broadcast agent advertisements at regular intervals of time. If the mobile node needs to get a care-of address and does not wish to wait for the periodic

[6] A Foreign Network can be defined as any network that is not the home network.

[7] The concept of IP-within-IP encapsulation and tunneling is described in detail in later sections.

advertisement, the mobile node can broadcast or multicast a solicitation that will be answered by any foreign agent or home agent that receives it. Home agents use agent advertisements to make themselves known, even if they do not offer any care-of addresses.

An agent advertisement performs the following important functions:

- Allows for detection of mobility agents.
- Lists one or more available care-of addresses.
- Informs the mobile node about special features provided by foreign agents, for example, alternative encapsulation techniques.
- Lets the mobile node determine the network number and status of their link to the Internet.
- Lets the mobile node know whether the agent is a home agent, foreign agent, or both, and therefore whether it is on its home network or a foreign network.

Mobile nodes use router advertisements as defined in [6.SED.1991] to detect any changes in the state of mobility agents available at the current point of attachment[8]. If advertisements are no longer detectable from a foreign agent that previously had offered a care-of address to the mobile node, the mobile node should presume that the foreign agent is no longer within range of the mobile node's network interface. In this situation, the mobile node should begin to hunt for a new care-of address, or possibly use a care-of address known from advertisements it is still receiving. The mobile node may choose to wait for another advertisement if it has not received any recently advertised care-of address, or it may send an agent solicitation.

Due to the unavailability of an Internet key management protocol, agent discovery messages are not required to be authenticated [7.JS.1996].

2.2 Registering the care-of address

Once a mobile node has a care-of address, its home agent must find out about it. The process begins when the mobile node, possibly with the assistance of a foreign agent, sends a registration request with the care-of address information. When the home agent receives this request, it adds the necessary information to its routing table, approves the request, and sends a registration reply back to the mobile node. Although the home agent is not required by the Mobile IP protocol to handle registration requests by updating entries in its routing table, doing so offers a natural implementation strategy, and thus necessary.

Registration requests contains parameters and flags that characterize the tunnel[9] through which the home agent will deliver packets to the care-of address. When a home agent accepts the request, it begins to associate the home address of the mobile node with the care-of address, and maintains its association until the registration lifetime[10] expires. The triplet that contains the home address, care-of address and the registration lifetime is called a binding for the mobile node. A registration request can be considered a binding update[11] sent by the mobile node.

A binding update is an example of a remote redirect, because it is sent remotely to the home agent to affect the home agent's routing table. This makes the need for authentication very clear. The home agent must be certain that the registration was originated by the mobile node and not by some other malicious node pretending to be the mobile node. A malicious node could cause the home agent to alter its routing table with erroneous care-of address information, and the mobile node would be unreachable to all incoming communications from the Internet.

The need to authenticate registration information has played a major role in determining the acceptable design parameters for Mobile IP. A detailed discussion of authentication is beyond the scope of this paper, however it is worthwhile to briefly mention how identification fields in the registration message are modified to improve authentication.

Mobile IP includes within the registration message a special identification field that changes with every new registration; this is mandated within Mobile IP. There are two main ways to make the identification field unique. One is to use a timestamp; then each new registration will have a later timestamp and thus differ from previous

[8] This is known as *agent solicitation.*

[9] A Tunnel is the path taken by encapsulated packets. It is the path that leads packets from the home agent to the foreign agent. Tunnels will be explained in more detail in later sections.

[10] Registration lifetime is how long the mobility agents may use the binding.

[11] A binding update is a message that supplies a new binding to an entity that needs to know the new care-of address for a mobile node. The binding update contains the mobile node's home address, instead of being one offered by a foreign agent.

registrations [7.JS.1996]. The other is to cause the identification to be a pseudorandom number. With enough bits of randomness, it is highly unlikely that two independently chosen values for the identification field will be the same [1.CP.2002].

This identification field is also used by the foreign agent to match pending registration requests to registration replies when they arrive at the home agent and to subsequently be able to relay the reply to the mobile node. The foreign agent also stores other information for pending registrations, including the mobile node's home address, the mobile node's Media Access Layer (MAC) address, the source port number for the registration request from the mobile node, the registration lifetime proposed by the mobile node, and the home agent's address. The foreign agent can limit registration lifetimes to a configurable value that it puts into its agent advertisements. The home agent can reduce the registration lifetime, which it includes as part of the registration reply, but it can never increase it [1.CP.2002].

2.3 Tunneling to the care-of address

The default encapsulation mechanism that must be supported by all mobility agents using Mobile IP is IP-within-IP (IPIP) [8.CP.1996.2]. Using IPIP, the home agent, which is the tunnel source, inserts a new IP header, or tunnel header, in front of the IP header of any datagram addressed to the mobile node's home address. The new tunnel header uses the mobile node's care-of address as the destination IP address, or tunnel destination. The tunnel source IP address is the home agent, and the tunnel header uses 4 as the higher-level protocol number, indicating that the next protocol header is again an IP header. In IPIP the entire original IP header is preserved as the first part of the payload of the tunnel header. Intermediate routers are unaware of this encapsulation.

In order to recover the original packet, the foreign agent merely has to eliminate the tunnel header and deliver the rest to the mobile node. In order to reduce header overhead, minimal encapsulation can be used instead of IPIP [9.CP.1996.3], however, this increases the complexity of processing the header as some of the information from the tunnel header is combined with information in the inner encapsulation header [10.CW, et al.2002].

3. EFFECTIVE USE OF MOBILE IP

As can be gathered from reading the previous two sections, the uses of Mobile IP are numerous. They can range from simple Internet access from handheld devices to allowing interactive sessions between employees in different locations while any one or all are on the move.

This is perhaps best demonstrated though a rudimentary example – imagine a scenario where an employee unplugs a mobile computer form its dock in the office. The computer, Mobile IP enabled, will be able to continue downloads, conduct Voice over IP (VoIP) sessions, and continue any other networking task without any interruptions, even while the employee is on his or her way home and finally from home (hopefully, the employee will not be working from home too much). Connectivity will first be transferred from the office Local Area Network (LAN), then to a cellular network and finally to the connection to the Internet that the employee has from home.

Another effective use of Mobile IP is the concept of Mobile Networks. According to Cisco Systems [11.CS.2001]:

> Cisco Mobile Networks enables a router and its subnets to be mobile while continuing to maintain IP connectivity transparent to the IP hosts connecting to the network through this mobile router. This solution enables the IP hosts on a LAN connected to such a mobile router to transparently connect to the parent network while the LAN is in motion. The mobile router ensures network connectivity from a mobile environment.
>
> The Mobile Networks solution enables entire networks to roam. For example, this enables a plane to fly around the world while passengers stay connected to the Internet. In this case, each passenger's IP device is a node on the mobile network connected to a router on the plane. The nodes on such a mobile network are not aware of any IP mobility at all. The Mobile Networks solution that is running on the router in the plane 'hides' the IP roaming from the local IP nodes so that the local nodes appear to be directly attached to the home network.

These are just some of the examples of the effective uses of Mobile IP. However, Mobile IP is expected to become even more important as wireless networks and IPv6[12] become popular. Cellular vendors are pushing Mobile IP hard as a way to allow seamless roaming between Third-Generation (3G) networks and higher bandwidth hot spots based on Bluetooth or Wi-Fi (802.11b).

[12] There is a brief discussion of IPv6 implementations in Mobile IP in section 6 of this paper.

To make clear some of the underlying technology that aid the effective use of Mobile IP, some of the concepts explained in the previous section will be reiterated with effective implementation kept in mind.

As explained, every kind of Mobile IP gives the mobile node two IP addresses, a permanent address on its home network (home address), and a care-of address (required if the mobile node is on a foreign network). The home address is the one that higher-level protocols use, while the care-of address signifies the node's actual location within a network and its subnets.

Whenever the mobile node moves to a new network, it must acquire a new care-of address on the network it is visiting. In IPv4 this means requesting one from a special mobility agent known as the foreign agent. Most likely, this agent will essentially be a Dynamic Host Configuration Protocol (DHCP) server. This server should have some Authentication, Authorization, and Accounting (AAA) functionality added on to the foreign network. AAA is basically an infrastructure, which allows "trust" in communications [12.SN, RP.2002].

Remote Authentication Dial In User Service (RADIUS) is the dominating AAA protocol in IP networks today [13.CR, et al.2000], while Diameter is the successor of this well-known protocol. Diameter is the currently standardized within the IETF AAA working group. Both RADIUS and Diameter are flexible and extensible protocols. Additionally Diameter is built to be backwards compatible with RADIUS. Both RADIUS and Diameter protocols are and will be used for roaming Mobile IP users [14.PC, et al.2002].

Back at the home network, another mobility agent, the home agent, usually an edge router with some AAA functions keeps track of all the mobile nodes with permanent addresses on that network, associating each with its care-of address. The mobile node keeps the home agent informed of its whereabouts by sending a binding update via ICMP, whenever its care-of address changes. These updates can incorporate a digital certificate, to ensure that they are actually sent by the mobile agent, rather than an attacker seeking to impersonate him or her.

When information (packets) is to be sent to the mobile node, the home agent intercepts these packets and sends them to the mobile node at its care-of address via tunneling. This allows the sender to send packets to a permanent IP address (which belongs to the home network).

In IPv4, this is the simplest way to ensure mobility, but this adds extra routing hops that tend to use more bandwidth and increase latency. The possibility of increased latency is of particular concern for wireless networks where preventing latency is paramount. In the original version of mobile IPv4, standardized in 1996, mobile nodes were supposed to send replies directly to senders, for compatibility with higher-level protocols, the "source" address field in these packets had to be the permanent address on the home network, even though routers on the Internet would see that the packets were actually coming from the care-of address on the visited network. This is a serious problem now.

To curb Denial of Service (DoS) attacks, where malicious packets often claim to be from fake IP addresses, routers on the Internet began to incorporate ingress and outgress filtering [15.PF, et al.1998]. Routers would only allow a packet through if its source address field was consistent with its origin. To get around these filters, mobile IPv4 was updated in 2002 to include reverse tunneling [16.GM.2001]. Instead of taking a triangular path (Figure 3.1), all packets travel via the home network only (Figure 3.2); this is known as Reserve Tunneling. Reverse Tunneling wastes bandwidth and adds to latency.

Figure 3.1
Triangular Routing

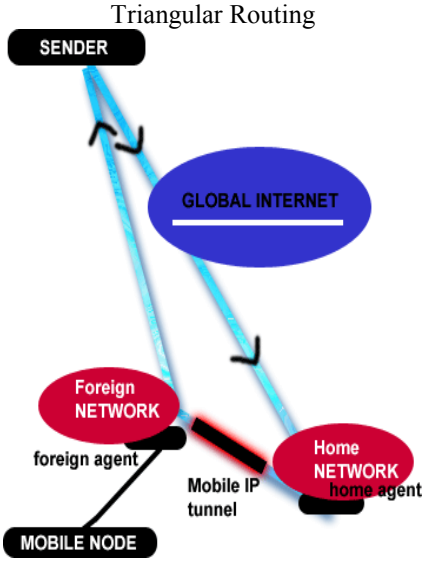
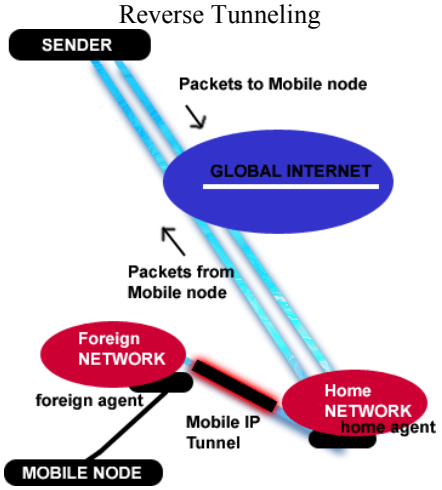


FIGURE 3.2
Reverse Tunneling



This problem is easily solved in Mobile IPv6 by avoiding tunneling as much as possible. Though the first few packets of every session are still tunneled via the home agent, the mobile node also sends binding updates to every correspondent. Future packets can be sent directly, just as if the mobile node belonged on the network it was visiting. Mobile IPv6 can accomplish this by using extensible headers – a feature allowing IPv6 packets to contain extra protocol information to deal with issues such as Quality of Service (QoS) and prioritization. The header of each packet can contain both the home and care-of address, satisfying both higher-level protocols and Internet routers. Section 6 briefly outlines these and other benefits of Mobile IPv6.

4. PROTOCOLS OF MOBILE IP

Mobile IP is a large part, but by no means the only part of mobile computing and networking. A good understanding of Mobile IP involves studying the relationships between the various layers of network protocols. Each layer should present a clear model of operation to the architect. Once the model is identified, the effects of mobility can be studied in relation to it.

The Internet networking stack showing common protocols is shown in the table below.

TABLE 4.1

Layers	**Common Protocols**
Application Presentation Session	HTTP, DNS, FTP, etc
Transport	TCP, UDP, RTP
Network	IP, ICMP, Mobile IP, etc
Data Link	IEEE 802.*, PPP, etc
Physical	Network Adaptor

Table 4.1 shows a simplified view of the International Standards Organization's (ISO) protocol stack as it applies to Internet networking. The major goal of Mobile IP protocol design was to handle mobility at the network layer and leave the other higher protocols unaffected. This allows for the existing routing infrastructure to be unaffected. This is also a big advantage to current applications, as they do not have to incorporate any new technologies either.

The cell[13] of a mobile node is the geographical location around the network (which supports mobile nodes). This area is defined by the propagation characteristics of the electromagnetic waves at the operating frequency. Infrared communications, for example, stop at the walls of a room, while radio frequency communications have much more complicated propagation characteristics. The ability of the hardware of the mobile host to detect signal strength and report it to the device driver will be helpful in realizing that the host is at the outer reaches of a cell and that it should try to switch to a different one. In addition, in the case of spread-spectrum of multi-channel radio communications, the ability to receive on multiple channels at once is highly desirable [17.JI, et al.1996].

The data link layer is responsible for link establishment and maintenance. Thus, physical effects from mobility are likely to require changes in this layer. Changes in position affect the Signal-to-Interference Ratio (SIR). Link layers that adapt forward error correction to SIR can exhibit variable bandwidth but far fewer lost packets. Wireless media typically introduce many other design requirements at the data link layer, such as encryption (to maintain a degree of confidentiality) and compression techniques (to fit more data in lower bandwidth) [18.TIA.1995].

An interesting problem to be dealt with in this layer is modifying the Address Resolution Protocol (ARP) to allow for mobility. In the data link layer, we either map a logical address to a physical address (non-mobile) or dynamically acquire an IP address for a moving node. ARP works by

[13] A cell is the geographic area in which a wireless device may function as desired. A large geographical area may be divided into cells in order to provide maximum signal strength to a wireless device.

broadcasting a request on the communication network with the IP address of the target in the packet. All hosts on the network will receive this packet, but only the node with the required IP address will respond[14].

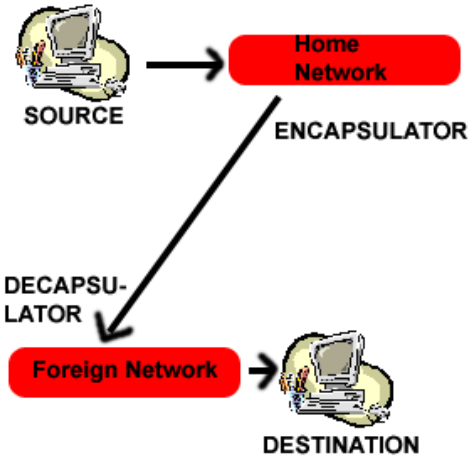
The problem with ARP and mobile nodes arises when one mobile node tries to get the address of another mobile node. If both the mobile nodes are on the same home network then ARP will work in the normal way. However, if one of the mobile nodes (MN1) is on another network, then the sending mobile node (MN2) will have to 'proxy-ARP' the request for MN1. MN2 will send all traffic intended for MN1 to its own home network, which will then encapsulate the request in an IP-within-IP datagram. The home network will then query all other networks in order to find out which one is handling MN1. This process continues until success or failure [17.JI, et al.1996]. This is also known as "Gratuitous ARP".

At the network layer, IP determines a path through a loosely confederated association of independent network links, routing from one network to another, while offering some services such as fragmentation, reassembly and checksumming. Hence, logically, changing the point of attachment (mobility) requires changing the routing of datagrams intended to and from the mobile node. This is done by encapsulation and decapsulation techniques. The standard technique for encapsulation and decapsulation is IP-within-IP (IPIP); this must be supported by default by all mobility agents.

Previously in this paper, IPIP has been used repeatedly without actually defining the concept in some detail. The concept of IPIP means that the payload of an IP datagram will actually contain another IP address and a payload. This is done to allow for delivery to an intermediate destination that would otherwise not be selected by the (network part of the) IP destination address field in the original IP header [8.CP.1996.2]. Once the encapsulated datagram arrives at this intermediate destination node, it is decapsulated, yielding the original IP datagram, which is then delivered to the destination indicated by the originally encapsulated IP datagram's destination address field.

This use of encapsulation and decapsulation of a datagram is referred to as tunneling, and the encapsulator and decapsulator are considered to be the endpoints of the tunnel shown in figures 3.1 and 3.2. The working for an encapsulator and decapsulator is shown in Figure 4.1. The end-points of a tunnel can also be thought of as the route between the home agent and the care-of address, handled by the foreign agent. It should be noted that encapsulation/decapsulation is employed only if the mobile node is on a foreign network.

FIGURE 4.1
Simple diagram showing IPIP scheme



SOURCE, shown in Figure 4,1 may be an Internet host or another mobile node; DESTINATION is the mobile node for which the datagram is intended. The home agent handles the start point of the tunnel where the encapsulation is done. At the end point of the tunnel, resides the foreign agent, which decapsulates the packet and delivers the packet to the mobile node. A problem arises at the decapsulation point. When the encapsulated packet is decapsulated the foreign agent sees that the destination field of the packet actually specifies the address of the home agent (home address). To avoid sending the packet right back to the home agent, the foreign agent must be smart enough to forward the packet to the intended mobile node identified by its hardware address, rather than the destination IP address.

Some problems with this kind of source routing[15] are [8.CP1996.2]:

[14] This fact is debatable, there are some cases when this might fail, or some rouge node may fake the IP address and respond instead. There are ways to prevent this, but they are beyond the scope of this paper.

[15] Source routing is a technique by which the sender can specify the route the datagram will take. Mobile IP utilizes Loose Source Record Route (LSRR), in which the sender gives one or more hops that the packet must go through. Another type, Strict source routing is not very common.

- Some current routers do not handle datagram modification or even source routing very well.
- There are some security issues with source routing.
- Encapsulation cannot be used if it is known that the tunnel end-point does not support it (i.e. it cannot decapsulate it).
- Firewalls may cause problems.

At the transport layer, TCP and other transport protocols attempt to offer a more convenient abstraction for data services than the characteristically chaotic stream of data emanating from IP. The vagaries and time dependencies of routers and Internet congestion often cause datagrams to be delivered out of order, duplicated or even dropped entirely before reaching their destination. TCP attempts to solve these and other problems, but offers little help in supplying a steady (constant bandwidth) stream of data.

Over time, TCP has been modified to treat dropped packets as an indication of network congestion, and therefore to throttle transmissions as soon as a lost packet is detected (by managing sequence numbers). In wireless communications, a lost or corrupted packet may be caused by noisy wireless channel, immediate retransmission is the best way to correct this, rather than delayed transmission. There is still a lot of research going on to find the best solution.

At the top layer are the application protocols. Depending on the transport model employed, application protocols are largely freed from the hassles of error correction, retransmission, flow control, and the like. However, mobility creates new needs at the application layer, which require additional protocol support: automatic configuration, service, discovery, link awareness, and environment awareness.

Essentially, these protocol support mechanisms form a set of middleware services. For example, a mobile computer might need to be configured differently at each different point of attachment. Among other things, a new DNS server, IP address, Link Minimum Transmission Unit (MTU), and default router may be required. These configuration items are usually thought of as being worked out at setup time for desktop systems, but for mobile computers no single answer can be sufficient. Recent deployment of DHCP goes some way toward resolving configuration difficulties, but is not the whole answer. Discovering services can be modeled as a requirement for automatic configuration, but is more naturally useful when services are located upon demand and according to the needs of application protocols. This need is being investigated by many means, Service Location Protocol being one of the primary ones.

The Service Location Protocol provides a scalable framework for the discovery and selection of network services. Using this protocol, computers using the Internet no longer need so much static configuration of network services for network-based applications. This is especially important as computers become more portable, and users less tolerant or able to fulfill the demands of network system administration [19.JV, et al.1997].

One of the challenges of such architecture of middleware is that it offers applications the opportunity to detect the state of the physical link, which may change dynamically and in turn affect the application's proper operation. Mobility requires that applications adapt to changing connection parameters such as bandwidth, error rate, and Round-Trip Times (RTT).

The following table concludes this section and summarizes what we know so far of the services provided/required by Mobile IP by showing the Network Stack and its corresponding nomadic services [20.CP.1998].

TABLE 4.2

NETWORK STACK	NOMADIC SERVICES
Application Layer	• Resource Discovery • Link Adaptation Layer
Transport Layer	• Congestion Control • Flow Control • Quality of Service
Network Layer	• Addressing • Routing • Location Management • Authentication
Physical/Link Layer	• Signal Modulation • Encryption • Compression • Interference • Channel Access/Selection

5. NEED FOR MOBILE IP

A good analogy for Mobile IP is the development of Mobile Cellular phones over the past decade. A lot of customers have migrated from the traditional fixed-point telephones (phones which require a fixed phone-jack in the wall) to Cellular phones which allow the user to be in motion while talking to someone. This mobility is achieved by using

technology that allows a user to maintain a constant connection to the receiver, and vice-versa.

A big difference in this analogy is the fact that it took a long time for mobile phones to become accepted in the consumer market. The prohibitive cost of cellular technology and limited areas of coverage discouraged many users from using cellular phones. Mobile computing is guaranteed to become a hot technology, as acceptance of mobile devices that can be modified to support Internet access is already popular with consumers. Some manufactures are already providing portable devices that allow wireless Internet access. Having already found user acceptance, mobile computing may become popular much faster.

As mentioned, one of the biggest advantages of Mobile IP is that the user can seamlessly move about without having to reconnect/reconfigure at every point of attachment. The main aim of Mobile IP with the aid of the protocols discussed in the previous section is to provide completely automatic, non-interactive reconnection to network activities from any point of attachment.

The important advantages of Mobile IP are:

- Location independent access to computing resources
- Wireless network access
- Ease of and comfortable operation – user can work from almost anywhere
- Economy – Mobile networking environments can save a lot of money by avoiding cabling costs and the subsequent maintenance (wear and tear) that come with wires.
- Software reusability – Existing applications do not have to be modified in order to be able to access a network over Mobile IP.
- Continuous connectivity

Some examples of Mobile IP in action have already been given in section 3 of this paper. Some more examples can be given where the use of Mobile IP can be critical to success:

- Collaborative office environments: Employees can freely move about with mobile devices and be able to interact, share and discuss data on their devices by being able to meet on a personal level. Airbus Wichita has a collaborative engineering facility that may be able to use this effectively [21.JL.2003].
- Hospitals: Doctors can get immediate information on their mobile devices on their patients without having to go to their terminals – they can the needed information in the presence of the patient. Using this kind of access will be particularly useful to Emergency Medical Services. Such services are already being tested [22.EMS].
- Battlefield Operations: Wireless voice communication has been mainstay in the military for many decades now, but with the advent of new military technology the need is there to provide devices that can receive/relay important battlefield data to/from a central computer [23.JJ.2002].
- Campus/College environments: Campus environments will benefit greatly from the implementation of Mobility within the campus environment.

Thus, once problems of mobile computing are ironed out, almost any computing environment, which needs Internetworking, can benefit greatly from the implementation of Mobile IP.

6. IMPLEMENTATIONS OF MOBILE IP

There are numerous implementations of Mobile IP. Already there are devices which support Mobile IP, and this is only furthered by technologies for Mobile IP that have been extended to different Operating systems such as Linux, and also to IPv6. This section will include:

6.1 A short discussion of Mobile IPv6
6.2 Mobile IP under Linux
6.3 Devices developed and being developed which support Mobile IP

6.1 A short discussion of Mobile IPv6

IPv6 was developed to include enhancements from IPv4 and many new technologies that are missing from IPv4. Mobile IPv6 contains many features that were sorely missed in Mobile IPv4. Principally the mobility support for Internet devices in possible and standardized for both IP protocol versions, IPv4 and IPv6, but due to the enhanced functionality and later design of IPv6 some features concerning the mobility support have been integrated more efficiently in Mobile IPv6 when compared to Mobile IPv4.

Some important advantages of Mobile IPv6 and differences from IPv4 are [24.WF, FH.2000, 25.DJ, et al.2003]:

- A mobile node requires a new care-of address every time it changes its point of attachment. There is a possibility under IPv4 that availability of these addresses becomes a problem. Due to the huge number of addresses available under IPv6, assigning addresses will never be a problem.
- Using anycast address of IPv6 enables a node to send a packet to one out of several systems having this anycast address assigned to one of

their interfaces. Mobile IPv6 makes efficient use of this mechanism for the dynamic host discovery mechanism by sending a binding update to the home agent any cast address and getting a response from exactly one home agent. This is absent from IPv4.

- Employing Stateless Address Auto configuration and Neighbor Discovery mechanisms Mobile IPv6 neither needs DHCP nor foreign agents to configure care-of addresses. Using Neighbor Discovery, Mobile IPv6 does not need to use ARP; this improves overall robustness of the protocol.
- Mobile IPv6 can use IPSec (IP Security Protocol) for all security requirements, like authentication, data integrity protection, and replay protection. This is perhaps the biggest difference between Mobile IPv4 and Mobile IPv6.
- Route optimization is greatly improved and is a fundamental part of Mobile IPv6. Mobile IPv6 route optimization can operate securely even without pre-arranged security associations. It is expected that route optimization can be deployed on a global scale between all mobile nodes and hosts that send data to them.
- Mobile IPv6 has no problems with ingress filtering, this improves efficiency and the wasteful technique of Reverse Tunneling is not required.
- The movement detection mechanism in Mobile IPv6 provides bidirectional confirmation of a mobile node's ability to communicate with its router from its current location.
- Packets in Mobile IPv4 require encapsulation if the mobile node is away from home, packets in Mobile IPv6 are sent using a Mobile IPv6 routing header, doing away with the need for encapsulation/ decapsulation techniques. This reduces the amount of resulting overhead compared to Mobile IPv4.
- Managing tunnel "soft state" is unnecessary in Mobile IPv6.

Conversely, Mobile IPv6 does not solve all general problems related to the use of mobility or wireless networks, some important examples [25.DJ, et al.2003]:

- Mobile IPv6 does not handle links with partial connectivity or unidirectional connectivity.
- Access control on a link being visited by another mobile node is not handled.
- Assistance for adaptive applications is absent.
- Mobile routers are not inherently supported; these have to be designed by the vendor.

A full discussion of IPv6 is worthy of a whole new paper, however, in this introductory paper it is sufficient to say that Mobile IPv4 contains many enhancements over Mobile IPv4 and is definitely the way to go for future mobile devices.

6.2 Mobile IP under Linux

There is currently a lot of research and implementations of Mobile IP under Linux, just to name a few:

- Mosquito Net - http://mosquitonet.stanford.edu/mip/
- Dynamics – HUT Mobile IP - http://www.cs.hut.fi/Research/Dynamics/
- Lancaster Mobile IPv6 Package - http://www.cs-ipv6.lancs.ac.uk/ipv6/MobileIP/
- Portland State University - http://www.cs.pdx.edu/research/SMN/

As can be seen, there are quite a few implementation of Mobile IPv6 for Linux, one common denominator is that most focus is now shifting towards Mobile IPv6 due the built-in IPv6 functionality in Linux kernels available today (Versions 2.1.x).

6.3 Devices developed and being developed which support Mobile IP

Companies which manufacture devices such as laptops, PDA's and similar devices are showing a lot of interest in Mobile IP as a way to provide reliable and "always-on" connections to services such as web browsing and e-mail.

Earthlink, Palm/Handspring, Sony, and other companies are investing heavily to get the edge in developing a device that can support seamless mobility.

Some examples of vendors developing products specifically to support Mobile IP are IPAQ 5455 from Hewlett Packard, supported by T-mobile. The IPAQ retails for about $600-$700 and has a modified version of the Windows operating system. Some new the newer models of PDA's from Palm and Sony also have nifty wireless features.

Research is underway at the labs of major vendors to produce devices which are fully capable if wireless communication, of course, the infrastructure needs to be in place for true nomadic computing. In some of the major metropolitan cities, where such infrastructure exists wireless communications have proved to be priceless.

This was best demonstrated by one of the biggest tragedies to happen to the United States. In the aftermath of the September 11th terrorist attacks, hundreds of office workers, finding themselves

without an office, turned to wireless communications to carry out the most vital of business tasks. Cellular technology and mobile computing also allowed rescue and other operations to be carried out with some semblance of reliability in the midst of tremendous destruction [26.SM.2003].

7. CONCLUSION

As this introductory paper has shown, Mobile IP has great potential. Within the IETF, Mobile IP already has a large number of both RFCs and Drafts, which indicates that Mobile IP has been a well researched field and that a great deal of current research is ongoing as well (A search in IETF website, http://search.ietf.org/, revealed 12,086 documents).

Test results have given added confidence that the Mobile IP specification is sound, implement able, and of diverse interest throughout the Internet community. Mobile IP specification has also been easily interpreted by network protocol engineers and network programmers [18.TIA.1995].

It is possible that the deployment pace of Mobile IP will track that of IPv6, or that the requirements for supporting mobility in IPv6 nodes will give additional impetus to the deployment of both IPv6 and mobile networking. The increased user convenience and the reduced need for application awareness of mobility can be a major driving force for adoption. Since both IPv6 and Mobile IP have little direct effect on the protocol stack, application designers should find this to be an acceptable programming environment. Of course, everything depends heavily on the willingness of platform and router vendors to implement Mobile IP and/or IPv6. Indications are there are most major vendors are already dedicating a lot of time and money to this field which has enormous potential.

REFERENCES

[1.CP.2002] Charles E. Perkins. IP Mobility Support. http://www.ietf.org/rfc/ rfc3344.txt. August 2002

[2.IP.2003] ipUnplugged White Paper – Mobility and Mobile IP, Introduction. 2003

[3.AD.2002] Andy Dornan. Mobile IP.http://www.networkmagaine.com/article/G20020429S0013. June 2002.

[4.CP.1996] Charles E. Perkins. Mobile IP, Ad-Hoc Networking and Nomadicity. http://citeseer.nj.nec.com/34824. html. 1996.

[5.CP.1999] Charles Perkins. Mobile Networking Through Mobile IP. http://www.computer.org/internet/v2n1/perkins.html. 1999.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

STUDY ON OUTLIERS AND HIERARCHICAL CLUSTERING ALGORITHMS USED FOR OUTLIER DETECTION

[1]Sandra Sagaya Mary.D.A. [2]Tamil Selvi.R [3]Suganya.R

[1 & 3] Research scholar, [2]Assistant Professor,
Computer Science Department Dr.SNS Rajalakshmi Arts and Science College
Coimbatore

ABSTRACT

Data Mining is the process of finding correlations or hidden patterns among the dozens of fields in a large database. There are many techniques available in data mining to find relations, among them clustering and classification are the two common techniques for finding hidden patterns. Here, we considered clustering which is most important technique for exploring data mining. Clustering is the process of grouping the related records based on their similarity and dissimilarity. In clustering, outlier is one of the major issue, that reduce accuracy of data. Outlier is an observation that is in an abnormal distance from other values in the dataset. Outlier occurs due to the changes that caused in system behavior, human error, poor data quality or by fraudulent behavior. This paper deals with the study of outliers, their types, applications and hierarchical clustering algorithms that deals with outlier detection.

Keywords: Outliers, Approaches, Applications, CURE, BIRCH, ROCK, CHAMELEON.

1. Introduction

Outliers are the data that are unusual from all the other data in the datasets. Though some data mining techniques deals with outlier detection, still there are problems arising due to outliers. Outlier detection is the important pre-processing step in data mining. Outlier detection is used in many fields like fraudulent detection, network intrusion, healthcare, industries and crime investigation. Outliers are divided into three main categories as supervised, semi-supervised and unsupervised. [1] In supervised outlier detection, the training data or labels will be available for both normal class and outlier class. This is similar to rare class mining. [2] In semi-supervised outlier detection, labels are available only for the normal data. This is more applicable than the supervised technique. [3] Unsupervised method does not need any training data. This is the technique that is most widely used. Accuracy of outlier detection lies in how well the clustering algorithms catch the cluster structure.

2. Types of Outliers

Outliers are detected by their nature. Outliers are categorized into:

- ✓ Point outliers
- ✓ Contextual outliers
- ✓ Collective outliers

2.1 Point outliers

Point outliers are the data which is considered as anomalous with respect to other data in a sample. This type of outlier is very simple and focused on many researches. In fraud detection point outlier detection is useful.

2.2 Contextual Outliers

Contextual outliers are individual data that is anomalous within a context and it requires notion of the context. In this type, every anomalous data have contextual attributes in which context of that data is described and behavioral attributes in which non-contextual attributes of data are provided. This outlier is mostly explored in spatial and time-series data in real world.

2.3 Collective Outliers

If collection of related data is anomalous with other data in dataset it is known as collective outliers. But in individual, they are not the outliers. This type of outlier is explored for spatial, sequence and graph data. This outlier will occur only in related datasets.

3. Outlier Detection Methods

There are many methods to detect outliers. Every techniques have their merits and demerits in their detection. Some of the approaches are:

- ✓ Distance Based Approach
- ✓ Distribution Based Approach
- ✓ Density Based Approach
- ✓ Clustering Based Approach
- ✓ Angle Based Approach

3.1 Distance Based Approach

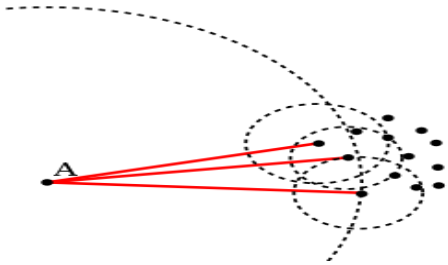
Distance based approach detects the outliers, based on their distance from the center of the cluster or centroid using the threshold value given by user. If the cluster has the value greater than the threshold value, then that cluster is said to be a outlier. Threshold value is the maximum distance of a data from the center of the cluster. The difficulty in this approach is that it requires threshold value by user which is difficult and it does not provide any ranking for outliers.

3.2 Distribution Based Approach

Distribution based approach is based on the statistical model. This method applies the statistical test for the data objects and the object with low probability to that test will be considered as the outlier. Distribution approach is possible only for the univariate data and does not suits high dimensional datasets. So, this method is less applicable in real world data.

3.3 Density Based Approach

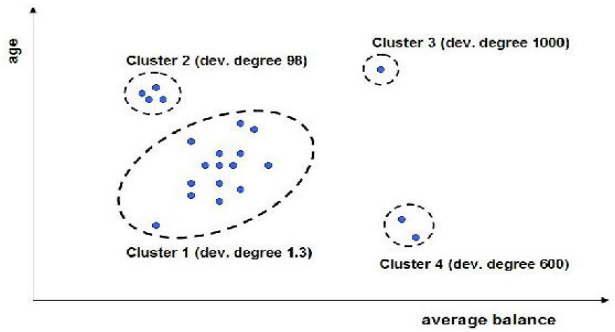
Density based approach is a method for finding outliers in spatial data. This detects outliers based on the local density of the neighbor. This approach has space-based and graph-based methods in which it considers both the attribute values and spatial relationship between the objects. In space based method, the definition is based on spatial neighborhood while in graph based method, definition is based on connectivity of graph. In density method, LOF(Local Outlier Factor) is used to detect local density.



Density based approach

3.4 Clustering Based Approach

In clustering based approach, first the data are formed into clusters using some clustering algorithms. Then the point is selected from the less dense clusters and the distance calculated between the denser cluster and point. If the point is far away from all other clusters, then it is declared as outlier. This approach is used when normal attributes are more than the abnormal attributes.



3.5 Angle Based Approach

In high dimensional data, angle based approach is much better than the distance based approach. The point is said to be an outlier, when its angle to other data points are small. Because the small distance in angle denotes that all other points are in some certain direction. If the spectrum angles are broad, then definitely that point will be surrounded by points in all possible directions.

4. Outlier Detection Applications

Outlier detection has several applications in real world such as intrusion detection, fraudulent detection, healthcare, industrial damage detection, image processing, loan application processing, monitoring.

Intrusion Detection- Intrusion detection refers to the unusual activities in the computer system. This can reduce the false alarm rate.

Fraudulent Detection- This refers to criminal activities in commercial organization. Outlier detection can be used for detection frauds in credit card, insurance claim, mobile phone and to detect the potentially problematic customers in loan application processing.

Healthcare Informatics- Current outlier detection techniques used to detect anomalous patients records, instruments faults and in time series data such as ECG and EEG.In pharmaceutical, it is used to detect the molecular structures.

Industrial Damage Detection-In industries, the data are extremely large and noisy. Outlier detection is used to detect intrusions in electronic systems, aircraft safety, abnormal energy consumption etc.

Defects in structure- In manufacturing, it finds the faulty production runs.

Image processing- monitors images by time and anomalous regions within the images. This is used in video surveillance, mammography etc.

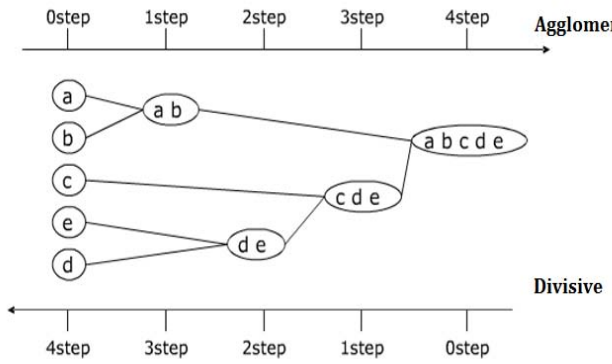
Detection in text- This detects novel topics or new stories in collection of documents.

Outlier detection also used in traffic monitoring, speech recognition, sensor networks, web applications etc.

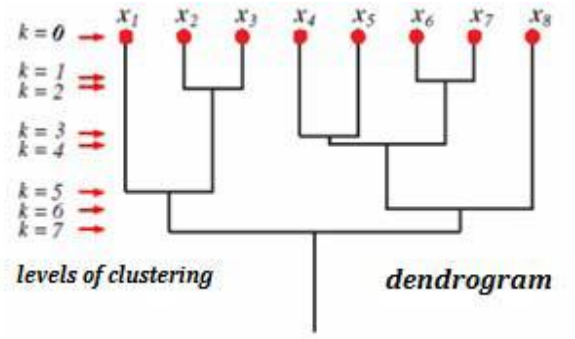
5. Hierarchical Clustering Algorithms for Outlier Detection

In clustering, hierarchical clustering is a technique that clusters the data objects in a hierarchical manner and are represented using a

dendrogram. This is a connectivity based clustering technique. Hierarchical clustering is easy to implement computationally and provide multi-solution clustering. This clustering is categorized into agglomerative and divisive clustering. Agglomerative clustering is the bottom-up approach in which it first initialize each object as single cluster and then find the distance or similarity between objects that are nearby. Finally merges the clusters that are similar, until all objects form single cluster. Divisive or deglomerative clustering is the top-down approach, which is the reverse of agglomerative clustering. This clustering divides the clusters into sub clusters from the single cluster until each cluster at lowest level is a cluster with single object. One disadvantage with this type of clustering is that once the clustering is performed, we cannot undo the previous process. Many algorithms developed in hierarchical clustering like CURE, BIRCH, CHAMELEON, ROCK, SLINK and CLINK which are also used in detecting outliers in the datasets.



Hierarchical clustering

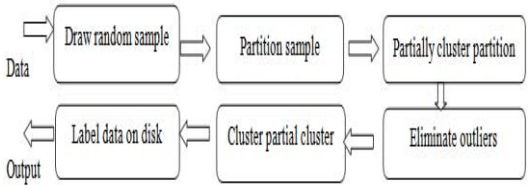


Dendrogram representation

6. CURE (Clustering Using Representatives)

Cure hierarchical algorithm was proposed by Guha, Rastogi and Shim in 1998. This algorithm is the combination of centroid approach and all-points approach which identifies the clusters by the representative points by choosing the well-scattered points and then shrinking them towards the center of the cluster. For large datasets, cure perform random sampling and partitioning to outperform all other existing algorithms without reducing clustering quality. This algorithm adjusts well to arbitrary shaped clusters. The procedure for cure is:

- First draw the random sample.
- Partition the sample
- Partially Cluster the partitioned sample
- Choose the representatives from the well scattered points
- Shrink them towards the center of the cluster
- The merging stops when there are only r clusters left. The r given by user.



CURE algorithm process

7. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

BIRCH algorithm is proposed by Zhang et al. in (1996) especially for large database. This minimize the number of I/O operations and provide the best quality clusters with single scan and available resources. Birch uses the CF tree which stores the clustering features of hierarchical clustering and this is the first algorithm to handle noisy data. It consists of four phases: (1) Loading, (2) Optional Condensing, (3) Global Clustering, and (4) Optional Refining. The complexity of the algorithm is O(n). The disadvantages of birch are they handle only numeric data and spherical shapes. The procedure is as follows:

- Scan the data and then CF tree is built
- While creating CF tree if memory exceeds, it builds a small tree from clustering feature and remaining objects. The above steps will be continued till the data fits the memory.
- Then using hierarchical method, it groups the cluster feature.
- Finally allocate them to nearest clusters.

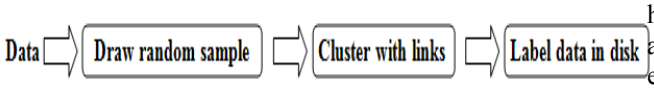
8. ROCK (RObust Clustering using linKs)

ROCK hierarchical clustering algorithm is fully based on links for handling large datasets. This algorithm generate good quality clusters and also scalable one. ROCK combines a conceptual point of view, nearest neighbor and hierarchical methods. In this algorithm, similarity is based on the number of points from different clusters that have their neighbors similar. The procedure is:

- Draw a random sample
- Link based hierarchical clustering method is applied to cluster the data points

- After merging, it considers each cluster as single data and keeps merging until there are no links remain between the clusters.

In rock algorithm, the new QROCK (Quick ROCK) algorithm has been developed for clustering categorical data easily. Their computational complexity is $O(n^2)$.



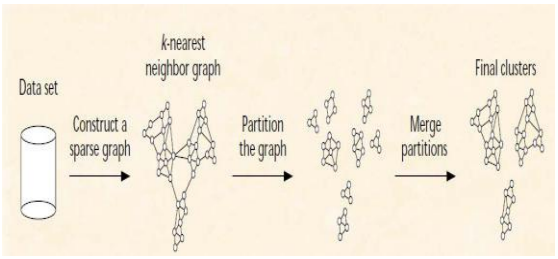
Process in ROCK

9. CHAMELEON

Chameleon is the hierarchical clustering algorithm based on dynamic modeling. This algorithm is applicable for arbitrary shapes, sizes and densities in two dimensional datasets. Chameleon works on both interconnectivity and closeness of the clusters. This algorithm measures the similarity based on the dynamic model and has two phases:

- First the data points are partitioned to form sub-clusters by graph partitioning.
- Then it merges all those sub-clusters by dynamic modeling to get final cluster.

The time complexity of chameleon is $O(n(\log_2 n+m))$.



Chameleon process

Hierarchical algorithms	**Model**	**Sensitivity to outliers**	**Time complexity**	**Space complexity**
CURE	Static	Less sensitive	$O(n^2 \log n)$	O(n)
BIRCH	Dynamic	Handle noise effectively	O(n)	-
ROCK	Static	-	$O(n^2+m_m m_a+n^2 \log n)$	$O(\min\{n^2 nm_m m_a\})$
CHAMELEON	Dynamic	-	$O(n(\log_2 n+m))$	-

Comparison of hierarchical clustering algorithms used for outlier detection

Conclusion

With this study, we have concluded that every algorithm has their own nature in clustering. In that BIRCH hierarchical clustering algorithm which is a dynamic model, handle noise very well and also have good time complexity, but they handle only clusters of spherical shapes. CURE algorithm handles arbitrary shaped clusters in an effective manner. Detecting Outliers play a vital part in clustering. Many algorithms have been developed to deal outliers. Additionally, the efficiency and effectiveness of a novel outlier detection algorithm can be defined as to handle large volume of data as well as high-dimensional features with acceptable time and storage, to detect outliers in different density regions, to show good data visualization and provide users with results that can simplify further analysis.

References

1. Yogita Rani and Dr. Harish Rohil, A study of hierarchical clustering algorithm, Volume 3, 2013.
2. Chris ding and Xiaofeng He (2002), Cluster Merging And Splitting In Hierarchical Clustering Algorithms.
3. S. Guha, R. Rastogi and K. Shim, CURE: An efficient clustering algorithm for large databases, Information Systems, 26 (2001), 35-58.
4. J. Harrington and M. Salibián-Barrera, Finding approximate solutions to combinatorial problems with very large data sets using BIRCH, Computational Statistics and Data Analysis, 54 (2010), 655-667.
5. P. Murugavel, Dr. M. Punithavalli," Improved Hybrid Clustering and Distance-based Technique for Outlier Removal", International Journal on Computer Science and Engineering, Volume 3, pp 333-339, 1 January 2011.
6. Silvia Cateni, Valentina Colla ,Marco Vannucci Scuola Superiore Sant Anna, Pisa," Outlier Detection Methods for Industrial Applications", ISBN 78-953-7619-16-9, pp. 472, October 2008
7. Shruti Aggrwal, Prabhdip kaur, "Survey of Partition Based Clustering Algorithm Used for Outlier Detection", June 2013.

The Publication of Conference Proceedings of National Level Conference on Research Issues in Computer Applications (NCRICA-14) held at Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.



International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0047
ISSN (Online): 2279-0055

International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

An Optimal Key Management in Wireless Network

[1] B Shanmugapriya, [2] S.Saranya

[1]M.Phil.(PT) Scholar, , [2]Assistant professor
Department of Computer Science, Sri Ramakrishna College of Arts and science for women, Coimbatore.

ABSTRACT

The development of several aggregation-based services, multiple multicast clusters are likely to synchronize in a single network, and users may contribute to multiple clustersconcurrently. The aim of actual Group Key Management is to secure a single group of element but It does not suitable for multicast group element, due to ineffective use of keys. We suggest a new GKM theme for multiple multicast teams, referred to as the master-key-encryption-based multiple cluster key management (MKE-MGKM) theme. The MKE-MGKM scheme exploits asymmetric keys, i.e., a base key and multiple derived keys, which are generated from the projected base key encryption (MKE) algorithm and iseffective distribution of the cluster key. It relieves the rekeying overhead by using the imbalance of the base and derived keys used for effective distribution of the cluster key. It relieves the rekeying overhead by using the asymmetry of the base and derived keys, i.e., if derived class is updated, other class should not change by only change the base key.

Keyword—Security, cluster key management, multicast, base key encryption

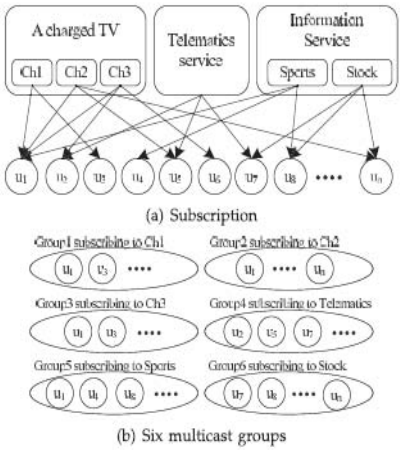
I INTRODUCTION

In the wireless network used one of the well-organized multicast methods. Multicast method used for transmitting data's from sender to the several destinations. Each and every transaction rates of efficiency should be received by sender. The multicast wireless network is increased the efficiency of group communication. These methods are called multicast service or multimedia broad cast. The broadcasting medium, however, makes the wireless network prone to numerous security attacks since anyone can simply snoop on messages transmitted within the air. To implement the multicast, i.e., the delivery of information solely to the members of a gaggle, in wireless networks, Access control mechanism for the broadcast method confidently to produce a accurate result. So it is successful optimal key management in wireless network. The way to produce an access control mechanism for secure group communication using symmetric key or group key. Group key only share a group members. Messages are converted to encrypted format and transmit to group members they are used group key and decrypt the message. This can provide secure group communication in wireless network. In this group key communication have some of the difficulties to provide an efficient for secure communication.

Group key should be updated in the particular interval time when the members are leaving and joining in the communication.

Since the existing group key management schemes have some of the limitation for using Rekeying in multicast service. In future the group of multicast network used in single network key management also support multicast group. Examples IEEE802.11 supports different multicast services eg., charged TV ,telematics service and information services.

In these services can be managed by several membership record of the service provided.if the subscription to a service is charged for either each channel (Ch1, Ch2, and Ch3) or content(sports and stock), the service supplierought to manage the extra user teams (e.g., channel-based or content-based groups) for correct accounting like Fig. As a result, the management overhead stemming from rekeying willconsiderably increase owing tothe amount of such teams. we abilitya brand new multiple group key management (MGKM) theme, named the master-key-encryption-based MGKM (MKE-MGKM) scheme, which mayscale backthe rekeying overhead from managing multiple cluster keys. The key plan of the MKEMGKM is to useANunevenencodingtheme, called the passe-partoutencoding (MKE), to reinforce the rekeying performance by assuaging the re keying



Multiple groups in a Wireless Networks

The Main Contribution:

1. We present a master key management algorithmic rulethat makes and updates a master key and multiple slave keys. A message encrypted by the master keywill be decrypted by everyof various slave keys, and contrariwise.

2. We present a rekeying mechanism for multiple teams by introducing a MKE-based key graph having the master and slave keys.

II RELATED WORKS

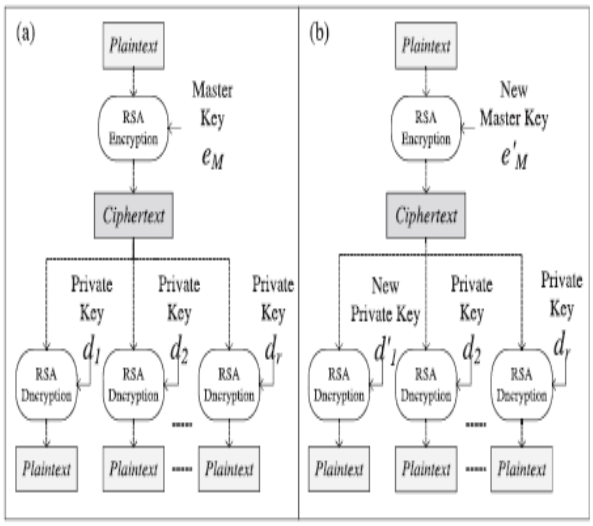
The KDC(Key Distribution Center) provide a new group key to all group members to invalidate old group key.so members should not allow leaving and joining to access in future message.This is called as forward or backward secrecy.

In a multicast group multiple members have shared a group key to encrypt/decrypt messages among themselves. After the member leaves, the old group key should be revoked and updated with a new group key. This rekeying process may cause a lot of key management overhead. Since the existing members do not have any shared secret keys except for the old group key

To solve these problem provide a new data structure called Logical Key Hierarchy (LKH). Group members shares Key Encryption Key(KEK),Traffic Encryption Key(TEK),Individual Key(IK). These keys are comprise Logical Key Trees, TEK is root node,IK is leaf node. The rekeying overhead is a logarithmic function of a group size. These tree based approach are one way function key and one way key derivation.

In the existing GKM scheme only used for single multicast group it cannot able to use multicast service in network. Multiple users can be access to multiple multicast groups. Each and every multicast groups can be access own group key management.All multicast groups independently used the rekeying procedure.

III MOTIVATION OF OUR APPROACH

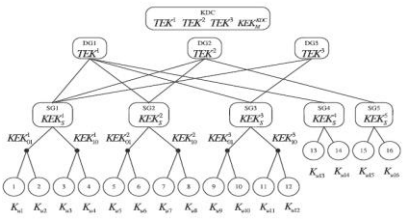


The Conceptual diagram of MKE

The existing HAC (Hierarchical access control) schemes can reduce the rekeying performance of the data structure is called hierarchical logical key tree. These schemes, KEK used to distributed TEK in efficient manner,KEK hierarchical key trees improve the rekeying originator from leaving user in symmetric TEK. These Schemes can be include another rekeying method in KEK also include a symmetric key.The single multicast group manage rekeying in TEK. The multiple multicast group due to consider several TEK and KEK can updated in multiple multicast group. Duplicate part of multiple multicast groups to be reduced in hierarchical IKG.It should not achieve a high rekeying performance in generic multicast environment in multicast services.

The MKE decrypt a cipher text each and every different key get the same plaintext, even one different key are repealed.

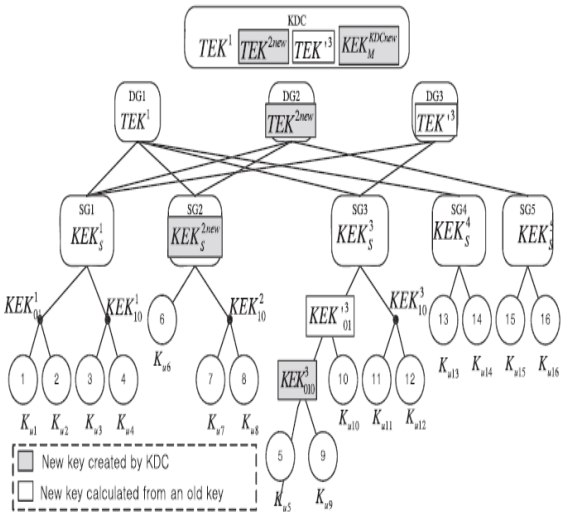
IV PROPOSED APPROACH FOR MGKM



An MKE-based key graph after the initial setup

In MBS is used for DG-Oriented encryption for ensure loe cost. Each and every encrypted data in TEK for all user corresponding DG share. Example Consider a broadcasting station providing 3 channels, every of that is for drama, sports, and news,severally. The three channels isthought to be3

MBSs if the content of every of the channels is encrypted and distributed severally. A user willsubscribe as several channels as desired, and hence, there isat the most3 DGs and seven SGs during this case.



An updated MKE-based key graph after rekeying

Step 1: Initial setup

1. Generating a master key and slave keys.
2. Constructing SG key tree.
3. Constructing an MKE based key graph.

Step 2: Rekeying at membership change

1. Revoking the old keys and generating a new master key.
2. Rekeying the new slave key and the KEKs inside the SG.
3. Broadcasting a new TEK encrypted with the new master

The KDC generates a master key, and as several public-private key pairs because the number of SGs, through the new projected master key management formula. the quantity of SGs grows exponentially because thevariety of DGs will increase. For example, allow us tothink abouttwenty DGs as 20 TV channels broadcasted by a supplier. If the provider uses all the potentialmixtures of channels.

This constructs as several SG key trees because thevariety of SGs. The leaf nodes of each SG key tree square measure the IKs of all users within the SG. This paper focuses on a way to manage multiple group keys, instead ofa way to generate ANeconomical key tree structure in an exceedinglycluster. thoughthere's no restriction on constructing the SG key trees, it is assumed that a binary tree structure is employed, to demonstrate the performance of the MGKM them.

The KDC finishes the initial setup by distributing the corresponding TEKs to the users in every SG in step withthe capability matrix. Since the initial distribution of the TEKs would be done just one occasion, it is not considered here however the TEKs square measure sent to the users of each SG at the initial setup; this may be done by using the user's IK or every SG's KEKs.

The analysis of MKE-MGKM scheme is reducing in terms of storage capacity in key information. These schemes employs in asymmetric encryption compare to symmetric it has provided more computational cost.

The storage overhead can be consider by memory capacity to maintaining the keys, It has directly access able to the number of keys to produce key sizes are same.

V CONCLUSION

An MGKM schemes has been enhance the performance of multiple group key used to hierarchy of the user or the data streams, other existing schemes used only symmetric keys,MKE-MGKM schemes used for asymmetric keys. It should be specify master key and multiple slave keys, It has generate by proposed master key algorithms.TEK can be produce efficient way of transaction. The number of rekeying message can be reduced for using MKE-MGKM schems.in these key management of MKE-MGKM schemes simpler to other schems.It has access less memory used for storing the keys compared to other schemes. MKE-MGKM schemes can be used for many practical solution for various group applications. It has required TV streaming service charged on a channel by channel basis.

VI REFERENCES

[1] IEEE Standard 802.16-2004, Part 16: Air Interface for Fixed Broadband Wireless Access Systems, IEEE, 2004.

[2] Third Generation Partnership Project, "Multimedia Broadcast/ Multicast Service; Stage 1 (Release 8)," Technical Specification

3GPP TS 22.146 v.8.3.0 (2007-06), June 2007.

[3] C.K. Wong, M.G. Gouda, and S.S. Lam, "Secure Group Communications Using key Graphs," ACM SIGCOMM Computer Comm.

Rev., vol. 28, pp. 68-79, 1998.

[4] D.M. Wallner, E.J. Harder, and R.C. Agee, "Key Management for Multicast: Issues and Architectures," IETF RFC 2627, http:// www.ietf.org/rfc/rfc2627.txt, June 1999.

[5] Y. Challal and H. Seba, "Group Key Management Protocols: A Novel Taxonomy," Int'l J. Information Technology, vol. 2, no. 1, pp. 105-118, 2005.

[6] S. McGrew, "Key Establishment in Large Dynamic Groups Using One-Way Function Trees," IEEE Trans. Software Eng., vol. 29, no. 5,pp. 444-458, May 2003.

[7] J.-C. Lin, F. Lai, and H.-C. Lee, "Efficient Group Key Management Protocol with One-Way Key Derivation," Proc. IEEE Conf. Local Computer Networks, pp. 336-343, http://doi.ieeecomputersociety. org/10.1109/LCN.2005.61, 2005.

[8] Y. Sun and K.J.R. Liu, "Hierarchical Group Access Control for Secure Multicast Communications," IEEE/ACM Trans. Networking,vol. 15, no. 6, pp. 1514-1526, Dec. 2007.

[9] Q. Zhang and Y. Wang, "A Centralized Key Management Scheme for Hierarchical Access Control," Proc. IEEE GLOBECOM, pp. 2067-2071, 2004.

[10] D. Wallner, E. Harder, and R. Agee, "Key Management for Multicast: Issues andArchitectures," IETF RFC 2627, 1999.

[11] R.L. Rivest, A. Shamir, and L. Adelman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," Comm. ACM, vol. 21, no. 2, pp. 120-126, 1978.

[12] K. Koyama, "A Master Key for the RSA Public-Key Cryptosystem," IEICE Trans. Information and Systems, pp. 163-170, 1982.